

**ARIZONA INDEPENDENT REDISTRICTING COMMISSION**

*Assorted Materials Opposing the U.S. Census Bureau's Use  
of Differential Privacy*

## Table of Contents

Pls.’ Motion for Prelim. Inj., <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.).....	1
Dec. of Dr. Michael Barber., <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.).....	75
Dec. of Thomas Bryan, <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.).....	104
Second Expert Report of Michael Barber, <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.) .....	198
Brief of Amicus Curiae Pr. Jane Bambauer, <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.) .....	212
Amicus Curiae Brief from Senate of Pennsylvania Republican Caucus et al., <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.) .....	243
Amicus Curiae Brief from the State of Utah et al., <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.).....	272
David A. Swanson et al., <i>The Effect of Differential Privacy Disclosure Avoidance System Proposed by the Census Bureau on 2020 Census Products: Four Case Studies of Census Blocks in Alaska</i> .....	315
David A. Swanson et al., <i>The Effect of Differential Privacy Disclosure Avoidance System Proposed by the Census Bureau on 2020 Census Products: Four Case Studies of Census Blocks in Mississippi</i> .....	320
Letter from Clark Benson, Politiadata LLC, to Steven Dillingham, Director, U.S. Bureau of the Census (Apr. 10, 2020) .....	337
Letter from K.C. Becker et al., Chair, Colorado Executive Committee of the Legislative Council, to Steven Dillingham, Director, U.S. Bureau of the Census (June 1, 2020) .....	345
David Van Riper, et al., <i>Feedback on the April 2021 Census Demonstration Files</i> (May 28, 2021) .....	351
Letter from Angela Hallowell, Main State Data Center, to Steven Dillingham, Director, U.S. Bureau of the Census (Feb. 20, 2020) .....	358
Letter from Marina Jenkins, National Redistricting Foundation, to Steven Dillingham, Director, U.S. Bureau of the Census (Apr. 24, 2020) .....	372

Policy Research Center, <i>2020 Census Disclosure Avoidance System: Potential Impacts on Tribal Nation Census Data</i> (May 2021) .....	375
Christopher Kenny et al., <i>The Impact of the U.S. Census Disclosure Avoidance System on Redistricting and Voting Rights Analysis</i> (May 28, 2021) .....	388
Jane Bambauer, et al., <i>Fool’s Gold: An Illustrated Critique of Differential Privacy</i> , 16 VAND. J. ENT. & TECH. L. 701 (Summer 2014) .....	406
Steven Ruggles, et al., <i>Differential Privacy and Census Data: Implications for Social and Economic Research</i> , AEA PAPERS AND PROCEEDINGS 2019, 109 (May 2019) .....	442
Letter from Meredith Gunter, Director, Weldon Cooper Center for Public Service, Univ. of Virginia, to Ralph Northam, Governor of Virginia (Jan. 23, 2020)...	448
Andrew Beveridge, <i>Controversial Census Bureau Plan that makes Data Less Accurate Goes to Court</i> , SOCIAL EXPLORER (May 2, 2021).....	451

**RECEIVED**  
**UNITED STATES DISTRICT COURT FOR THE**  
**MIDDLE DISTRICT OF ALABAMA**  
**EASTERN DIVISION**

2021 MAR 11 P 4:51  
THE STATE OF ALABAMA; ROBERT ADERHOLT, Representative for Alabama's 4th Congressional District, in his official and individual capacities; WILLIAM GREEN and CAMARAN WILLIAMS,

Plaintiffs,

v.

UNITED STATES DEPARTMENT OF COMMERCE; GINA RAIMONDO, in her official capacity as Secretary of Commerce; UNITED STATES BUREAU OF THE CENSUS, an agency within the United States Department of Commerce; and RON JARMIN, in his official capacity as Acting Director of the U.S. Census Bureau,

Defendants.

CIVIL ACTION NO. 3:21-CV-211

**THREE-JUDGE COURT REQUESTED  
PURSUANT TO 28 U.S.C. § 2284**

**PLAINTIFFS' MOTION FOR A PRELIMINARY INJUNCTION, PETITION FOR A WRIT OF MANDAMUS, AND MEMORANDUM IN SUPPORT**

Plaintiffs hereby move under Federal Rule of Civil Procedure 65 for an order preliminarily enjoining the defendants from both implementing differential privacy and enforcing the "February 12 Decision" to delay the provision of redistricting data. Plaintiffs additionally, and in the alternative, petition under 28 U.S.C. § 1361 for a writ of mandamus ordering the Secretary to comply with her statutory obligation to provide redistricting data under 13 U.S.C. §141(c) by March 31, 2021. Plaintiffs offer the memorandum incorporated below in support of their motion and petition.

Because of the impending statutory deadline, Plaintiffs request the Court set a briefing and hearing schedule and propose the following schedule: Defendants' Response due by March 23, 2021; Plaintiffs' Reply by March 26, 2021; and a hearing on this motion on March 29, 2021.

**TABLE OF CONTENTS**

TABLE OF CONTENTS..... i

TABLE OF AUTHORITIES ..... iii

INTRODUCTION ..... 1

BACKGROUND ..... 4

    A. Congress Requires Defendants to Provide “Tabulations of Population” for States to Use  
    for Redistricting ..... 4

    B. Alabama Relies on the Census Bureau’s “Tabulations of Population” to be Accurately  
    and Timely Reported..... 6

    C. The Census Bureau Adopts a Statistical Method Called “Differential Privacy” That  
    Will Cause the “Tabulations of Population” Used for Redistricting to be Inaccurate..... 9

        1. The Census Bureau Has Relied on Various Disclosure Avoidance Methods  
        to Successfully Protect the Privacy of Census Respondents in the Past..... 9

        2. The Census Bureau Adopts “Differential Privacy” for the 2020 Census ..... 12

        3. Differential Privacy is Unnecessary and Unproven in the Census Context..... 15

        4. Differential Privacy Will Result in Inaccurate Population Tabulations. .... 18

    D. The Bureau Delays Release of the “Tabulations of Population” ..... 24

    E. Plaintiffs Seek Relief From This Court. .... 25

JURISDICTION ..... 26

LEGAL STANDARD..... 26

ARGUMENT..... 26

    I. Plaintiffs Are Entitled To An Injunction Requiring Defendants To Provide Alabama With  
    Timely And Accurate “Tabulations Of Population” Unaffected By The Application Of  
    Differential Privacy..... 26

        A. Plaintiffs Are Likely To Prevail on the Merits. .... 29

1. The Application of Differential Privacy Violates the Census Act, Individual Plaintiffs’ Constitutional Rights, and Plaintiffs’ Rights Under Public Law No. 105-119, § 209.....	29
2. The Application of Differential Privacy Violates the Administrative Procedure Act.....	12
3. The February 12 Decision Violates the Census Act.....	43
4. The February 12 Decision Violates the Administrative Procedure Act .....	45
B. Without an Injunction, Plaintiffs Will Be Irreparably Harmed.....	50
1. The Inaccurate Population Tabulations Will Irreparably Harm Plaintiffs.....	50
2. The Delayed Population Tabulations Will Irreparably Harm Plaintiffs .....	55
C. The Benefits of an Injunction Far Outweigh the Costs.....	56
D. An Injunction Will Serve the Public Interest.....	57
II. In The Alternative, The Court Should Issue A Writ Of Mandamus.....	58
CONCLUSION.....	59
CERTIFICATE OF SERVICE .....	60

**TABLE OF AUTHORITIES**

**Cases**

*Arizona v. Inter Tribal Council of Az., Inc.*,  
570 U.S. 1 (2013)..... 31

*Azar v. Allina Health Servs.*,  
139 S. Ct. 1804(2019)..... 30

*Bennett v. Spear*,  
520 U.S. 154 (1997)..... 39, 40, 46

*Brown v. Thomson*,  
462 U.S. 835 (1983)..... 35

*Buckley v. Valeo*,  
424 U.S. 1 (1976)..... 35

*Cheney v. U.S. Dist. Ct.*,  
542 U.S. 367 (2004)..... 58

*Conn. Nat’l Bank v. Germain*,  
503 U.S. 249 (1992)..... 45

*Dep’t of Commerce v. U.S. House of Representatives*,  
525 U.S. 316 (1999)..... 4, 31

*Dep’t of Homeland Sec. v. Regents of the Univ. of Cal.*,  
140 S. Ct. 1891 (2020)..... 41, 47

*Edward J. DeBartolo Corp. v. Fla. Gulf Coast Bldg. & Const. Trades Council*,  
485 U.S. 568 (1988)..... 31

*Evenwel v. Abbott*,  
136 S. Ct. 1120 (2016)..... 7, 33

*Fed. Election Comm’n v. Akins*,  
524 U.S. 11 (1998)..... 33

*Fed. Election Comm’n v. Democratic Senatorial Campaign Comm.*,  
454 U.S. 27 (1981)..... 40

*Franklin v. Massachusetts*,  
505 U.S. 788 (1992)..... 39



*Georgia v. Evans*,  
316 U.S. 159 (1942)..... 37

*Heckler v. Ringer*,  
466 U.S. 602 (1984)..... 26, 58

*Karcher v. Daggett*,  
462 U.S. 725(1983)..... 33, 35, 37

*Kingdomware Technologies, Inc. v. United States*,  
136 S. Ct. 1969 (2016)..... 44

*Kirkpatrick v. Preisler*,  
394 U.S. 526 (1969)..... 33

*Lopez v. Davis*,  
531 U.S. 230 (2001)..... 44

*Maine Cmty. Health Options v. United States*,  
140 S. Ct. 1308 (2020)..... 44

*Maryland v. King*,  
133 S. Ct. 1 (2012)..... 55

*Merck Sharp & Dohme Corp. v. Albrecht*,  
139 S. Ct. 1668 (2019)..... 46

*Motor Vehicle Mfrs. Ass’n v. State Farm Mut. Auto. Ins. Co.*,  
463 U.S. 29 (1983)..... 41, 47

*Nat’l Urban League v. Coggins*,  
No. 5:20-CV-05799-LHK (N.D. Cal. Feb. 24, 2021)..... 48

*New York v. FERC*,  
535 U. S. 1 (2002)..... 46

*Ross v. Nat’l Urban League*,  
141 S. Ct. 18 (2020)..... 45

*Seymour v. Barabba*,  
559 F.2d 806 (D.C. Cir. 1977)..... 29

*Siegel v. Lepore*,  
234 F.3d 1163 (11th Cir. 2000) ..... 26, 56

<i>Smiley v. Holm</i> , 285 U.S. 355 (1932).....	6
<i>Susan B. Anthony List v. Driehaus</i> , 573 U.S. 149 (2014).....	35
<i>Thornburg v. Gingles</i> , 478 U.S. 30 (1986).....	34
<i>U.S. Army Corps of Eng’rs v. Hawkes Co.</i> , 136 S. Ct. 1807 (2016).....	39, 46
<i>U.S. House of Representatives v. U.S. Dep’t of Com.</i> , 11 F. Supp. 2d 76 (D.D.C. 1998).....	32
<i>United States v. Schmidt</i> , 675 F.3d 1164 (8th Cir. 2012) .....	38
<i>Utah v. Evans</i> , 536 U.S. 452 (2002).....	37
<i>Veith v. Pennsylvania</i> , 195 F. Supp. 2d 672 (M.D. Pa. 2002).....	22
<i>Wesberry v. Sanders</i> , 376 U.S. 1 (1964).....	35
<b>Alabama Statutes</b>	
Ala. Code § 17-5-7(b)(2) .....	9, 56
Ala. Code § 17-11-5(b).....	8, 9
Ala. Code § 17-11-12.....	8
Ala. Code § 17-13-5(a) .....	9
<b>Alabama Constitutional Provisions</b>	
Ala. Const. art. IV, § 47.....	9
Ala. Const. art. IX, §§ 197-200.....	7

**Rules**

Soliciting Feedback From Users on 2020 Census Data Products, 83 Fed. Reg. 34,111 (July 19, 2018) ..... 41

Federal Rule of Civil Procedure 65 ..... 3

**United States Constitutional Provisions**

U.S. Const. art I, § 2, cl. 3..... 4, 7; 31

U.S. Const. art. I, § 4, cl.1..... 6

U.S. Const., art. I, § 7..... 46

U.S. Const. amend. XIV, § 2 ..... 1, 4, 31

**United States Statutes**

5 U.S.C. § 706(2)(A), (B), (C)..... 39, 40, 45, 46

13 U.S.C. § 1..... 4

13 U.S.C. § 9..... 9, 41, 42, 43

13 U.S.C. § 141(a) ..... 4, 6

13 U.S.C. § 141(b)..... 5, 30

13 U.S.C. § 141(c) ..... passim

28 U.S.C. § 1331..... 26

28 U.S.C. § 1361..... 25, 58

28 U.S.C. § 2201(a) ..... 26

28 U.S.C. § 2284(a) ..... 25

**Acts of Congress**

Act of Dec. 23, 1975, Pub. L. No. 94-171, 89 Stat. 1023 (codified at 13 U.S.C. § 141(c))... 24, 44

Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies  
 Appropriations Act of 1998, Pub. L. No. 105-119, § 209, 111 Stat. 2440  
 (codified at 13 U.S.C. § 141 note) ..... passim

**Other Authorities**

Albert E. Fontenot, 2020 Census Update, Presentation to the Census Scientific Advisory  
 Committee March 18, 2021, <https://perma.cc/A4UM-FHCU> ..... 47

Amy Lauger et al., U.S. Census Bureau, *Disclosure Avoidance Techniques at the U.S. Census  
 Bureau: Current Practices and Research 2*  
 (Sept. 26, 2014), <https://perma.cc/2UXQ-SAFL> ..... 10

Andrew Reamer, *Counting for Dollars 2020: The Role of the Decennial Census in the  
 Geographic Distribution of Federal Funds*, Brief 7: Comprehensive Accounting of Census-  
 Guided Federal Spending: Part A: Nationwide Analysis (FY2017),  
<https://perma.cc/WQT9-DBYQ> ..... 52

Andrew Reamer, *Counting for Dollars 2020*, Brief 7: Comprehensive Accounting of Census-  
 Guided Federal Spending: Part B: State Estimates (FY2017),  
<https://perma.cc/8PWU-TM57> ..... 52

Antonin Scalia & Bryan A. Garner, *Reading Law: The Interpretation of Legal Texts* (2012) .... 38

*Census Bureau Statement on Redistricting Data Timeline*, U.S. Census Bureau (Feb. 12, 2021),  
<https://perma.cc/A2SZ-7L5Q> ..... 25

*Census Data Snafu Upends 2022 Elections*, Politico (Mar. 1, 2021), <https://perma.cc/DZ5N-275Y> ..... 7

Cynthia Dwork, *Differential Privacy: A Cryptographic Approach to Private Data Analysis*,  
*in Privacy, Big Data and the Public Good* (Julia Lane et al., eds., 2014) ..... 13

Harvard University Privacy Tools Project, *Differential Privacy*, <https://perma.cc/T7NJ-N397>  
 (last visited Mar. 2, 2021) ..... 13

JASON, *Formal Privacy Methods for the 2020 Census* (Apr. 2020),  
<https://perma.cc/G8ZM-YNN6> ..... 29

Jeff Zalesin, *Beyond the Adjustment Wars: Dealing With Uncertainty and  
 Bias in Redistricting Data*, 130 Yale L.J. Forum 186, 187-89 (2020) ..... 51

John M. Abowd, U.S. Census Bureau, Presentation to the 24th ACM SIGKDD Conference on  
 Knowledge Discovery and Data Mining, *The U.S. Census Bureau Adopts Differential  
 Privacy* (Aug. 23, 2018), <https://perma.cc/USZ6-ZPLC> ..... 12

John M. Abowd, U.S. Census Bureau, Presentation to the Am. Ass’n for the Advancement of Science, *Staring Down the Database Reconstruction Theorem* (Feb. 16, 2019), <https://perma.cc/P3YV-FXPG> ..... 16

Laura McKenna, U.S. Census Bureau, *Research & Methodology Directorate: Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Census of Population and Housing Public Use Microdata Samples* (Apr. 2019), <https://perma.cc/9LBN-5BWV> ..... 9, 10, 11

Laura Zayatz et al., U.S. Census Bureau, *Disclosure Avoidance for Census 2010 and American Community Survey Five-year Tabular Data Products* (Nov. 23, 2009), <https://perma.cc/GF4V-QTVA> ..... 11

Letter from JASON to U.S. Census Bureau at 5, fig. 1 (Feb. 8, 2021), <https://perma.cc/D3RF-TEBA> ..... 48

Merriam-Webster’s Collegiate Dictionary (10th ed. 1993)..... 30

Michael B. Hawes, U.S. Census Bureau, *Implementing Differential Privacy: Seven Lessons From the 2020 United States Census*, Harvard Data Science Review (Apr. 30, 2020), <https://perma.cc/DB66-9B5R>..... 18

Michael Hawes, U.S. Census Bureau, *Title 13, Differential Privacy, and the 2020 Decennial Census 22* (Nov. 13, 2019), <https://perma.cc/MRQ2-67WG>..... 13, 16, 17

Mike Mohrman, Letter to Steve Dillingham, Director, U.S. Census Bureau (Feb. 6, 2020), <https://perma.cc/MC3G-62PT>..... 21

Mike Mohrman, *The Challenge of Differential Privacy: Confidentiality vs. Usability* (Sept. 15, 2020), <https://perma.cc/4FA7-G4EF> ..... 20

Nat’l Conf. of State Legislatures, *Differential Privacy for Census Data Explained* (Feb. 1, 2021), <https://perma.cc/DA93-36GA> ..... 10, 15

Nat’l Redistricting Foundation, Letter to Steven Dillingham, Director, U.S. Census Bureau (Apr. 24, 2020), <https://perma.cc/3QK8-65VN> ..... 21

*Press Release: Census Bureau Statement on Redistricting Data Timeline*, U.S. Census Bureau (Feb. 12, 2021), <https://perma.cc/TY9T-UNDM>..... 24

*Random House College Dictionary 1337* (revised ed. 1975) ..... 29

Simson L. Garfinkel, John M. Abowd, & Sarah Powazek, *Issues Encountered Deploying Differential Privacy 3* (Sept. 6, 2018), <https://perma.cc/7TZQ-AFTD> ..... 14, 17

Simson L. Garfinkel, U.S. Census Bureau, *Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the 2018 End-to-End Test* (Sept. 15, 2017), <https://perma.cc/4J8B-ZEXM> ..... 12

Steven Ruggles et al., *Differential Privacy and Census Data: Implications for Social and Economic Research*, 109 AEA Papers and Proceedings (May 2019), <https://perma.cc/GW29-GNAV> ..... 12, 16, 17, 43

U.S. Census Bureau, 2010 Demonstration Data Products (rev. Apr. 16, 2020), <https://perma.cc/KK5M-KLRL>..... 18

U.S. Census Bureau, *2020 Census Operational Plan: A New Design for the 21st Century—Version 4.0* (Dec. 2018) ..... 13, 39

U.S. Census Bureau, *2020 Census Response Rate Update: 99.98% Complete Nationwide* (Oct. 19, 2020), <https://perma.cc/MFE3-8PDP> ..... 48

U.S. Census Bureau, *2020 Census State Redistricting Data (Public Law 94-171) Summary File 7-3* (Feb. 2021), <https://perma.cc/9HWC-492T> ..... 15

U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), <https://perma.cc/D6VJ-N5Z3> ..... 14, 18, 35, 49

U.S. Census Bureau, *2020: Census: Our Mission to Count Everyone* (Dec. 2020), <https://perma.cc/43R7-LNAL> ..... 1

U.S. Census Bureau, *Adapting Field Operations To Meet Unprecedented Challenges* (Mar. 1, 2021), <https://perma.cc/AU4S-9GXC>..... 48

U.S. Census Bureau, *Census Data Processing 101* (Feb. 11, 2020), <https://perma.cc/E8JK-4S9V>. ..... 48

U.S. Census Bureau, *U.S. Department of Commerce Secretary Wilbur Ross and U.S. Census Bureau Director Steven Dillingham Statement on 2020 Census Operational Adjustments Due to COVID-19* (Apr. 13, 2020), <https://perma.cc/C2RG-UXBX>..... 45

## INTRODUCTION

This case is about (1) the Census Bureau’s unprecedented decision to report skewed, inaccurate redistricting data to the States in place of the tabulations of population the Bureau is required by statute to provide, and (2) the Bureau’s decision to ignore its statutory deadline for reporting redistricting data.

Every ten years, the Census Bureau conducts the monumental task of “counting the whole number of persons in each State.” U.S. Const. amend. XIV, § 2. The decennial census’s importance is hard to overstate, as the tabulations of population the Bureau produces to the President and the States will shift political power between the States and within them, and will direct the flow of billions of dollars in federal and state funding. Thus, the Bureau’s mission is “to count everyone once, only once, and in the right place.”<sup>1</sup> And then the Bureau must report to States detailed “[t]abulations of population” at the sub-state level so States can draw new congressional, legislative, and other representative districts. 13 U.S.C. § 141(c). But with this census, for the first time ever, rather than provide States the actual results of the count, the Bureau intends to provide numbers produced by a still developing confidential algorithm. And in addition to abandoning its duty to provide true population data to the States, the Bureau has refused to produce redistricting data on time. Both decisions violate the law, harming Alabama and its residents. And both decisions should be immediately enjoined. *See* Complaint, Doc. 1.

*First*, the manipulated numbers. Congress has ordered the Secretary of Commerce to report to each State accurate “[t]abulations of population” for subparts of each State for use in “legislative apportionment or districting of such State.” 13 U.S.C. § 141(c). But the Secretary of Commerce, through the Census Bureau, has announced that she will instead provide the States purposefully

---

<sup>1</sup> *See* U.S. Census Bureau, *2020: Census: Our Mission to Count Everyone* (Dec. 2020), <https://perma.cc/43R7-LNAL>.

flawed population tabulations. The Bureau intends to use a statistical method called differential privacy to intentionally skew the population tabulations given to States to use for redistricting. Thus, the Bureau might “count everyone once,” and “only once,” but it won’t count them “in the right place.”<sup>2</sup> In fact, the *only* counts that will be unaltered by differential privacy will be the total population of each State, the total housing units at the census block level, and the number of group quarters facilities by type at the census block level. *All other tabulations*—including how many people live in a census block, town, or county—will be intentionally scrambled, denying Alabama accurate information about where Alabamians actually live.

Without relief from this Court, Plaintiffs will be irreparably harmed by this decision. It will violate Alabama’s right to receive lawfully composed population tabulations at the sub-state level, harm the State’s sovereign interest in drawing districts that provide its citizens fair representation, and increase the chance that Alabama will face litigation over its redistricting decisions. Relatedly, Representative Robert Aderholt, William Green, Camaran Williams, and others across the State will face a substantial risk that their voting power will be diluted when the Bureau purposefully misreports the number of people living in different areas of the State. That is why Congress has determined that the unlawful use of statistical methods to formulate redistricting data harms congressional representatives and the people whose representation could be affected. *See* Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act of 1998, Pub. L. No. 105-119, § 209(d), 111 Stat. 2440 (codified at 13 U.S.C. § 141 note).

*Second*, the unlawful delay. Not only does the Bureau intend to produce false redistricting numbers; it intends to produce numbers half a year behind schedule. Congress required the Bureau to engage in a five-year collaborative process with the States to ensure delivery of redistricting

---

<sup>2</sup> *See id.*



data by no later than March 31, 2021. *See* 13 U.S.C. § 141(c). Alabama upheld its end of the deal, but the Bureau has unilaterally decided that it will instead submit data to the States by September 30, 2021. The Bureau has no authority to grant itself this extension and deprive Alabama of information to which it is entitled. That is especially so because the Bureau's delay imposes substantial costs on Alabama as the State seeks to meet its constitutional obligations and run its 2022 statewide elections effectively and in accordance with State law.

Plaintiffs thus respectfully move this Court for a preliminary injunction under Federal Rule of Civil Procedure 65. They ask the Court to (1) enjoin Defendants from using differential privacy in connection with the 2020 census, and (2) enjoin Defendants from delaying the release of the redistricting data to the States. In the alternative, Plaintiffs petition the Court for a writ of mandamus under 28 U.S.C. § 1361 requiring Defendants to meet the statutory March 31 deadline for releasing the redistricting data.

Relief is necessary *now*. As Congress has recognized, “the decennial enumeration of the population is a complex and vast undertaking, and if such enumeration is conducted in a manner that does not comply with the requirements of the Constitution or laws of the United States, it would be impracticable for the States to obtain, and the courts of the United States to provide, meaningful relief after such enumeration has been conducted.” Pub. L. No. 105-119, § 209(a)(8). That is particularly true here. Depending on how differential privacy is implemented, the Census Bureau may argue that it will be impossible to unscramble the egg by ever delivering the accurate numbers without creating significant privacy risks from the release of two datasets. In that case, absent immediate action by this Court, the true population tabulations will never be known. On the other hand, even if corrected tabulations could one day be released, publishing the faulty numbers first will irreparably harm Plaintiffs. States like Alabama are already facing a time-crunch in

their redistricting schedules due to Defendants' delay. Redistricting will thus begin as soon as the Bureau delivers the population tabulations. If the Bureau then releases a second set of tabulations, States will be forced to scrap their redistricting plans and begin the process anew—or face a barrage of lawsuits for relying on the flawed tabulations. Either way, injunctive relief from this Court is needed to prevent these harms.

### **BACKGROUND**

#### **A. Congress Requires Defendants to Provide “Tabulations of Population” for States to Use for Redistricting.**

Under the Constitution, representation in the House of Representatives is “apportioned among the several States according to their respective numbers, counting the whole number of persons in each State, excluding Indians not taxed.” U.S. Const. amend XIV, § 2. There are two main components of this apportionment. The first is the division of congressional seats among the 50 States. The second is the redistricting process within each State that follows that division. *See Dep’t of Commerce v. U.S. House of Representatives*, 525 U.S. 316, 328-34 (1999) (discussing the “purposes” of apportionment).

To determine the “whole number of persons in each State,” an “actual Enumeration”—the decennial census—is required every ten years, “in such Manner as [Congress] shall by Law direct.” U.S. Const. art I, § 2, cl. 3. Congress enacted the Census Act to direct the “Manner” in which the decennial census occurs. *See generally* 13 U.S.C. § 1 *et seq.* Under the Act, the Secretary of Commerce, who oversees the U.S. Census Bureau, is required to, “in the year 1980 and every 10 years thereafter, take a decennial census of population as of the first day of April of such year.” 13 U.S.C. § 141(a).

Following the census, the Secretary has two primary sets of population numbers she must report. The first is “[t]he tabulation of total population by States” that is used for “the apportionment of Representatives in Congress among the several States.” *Id.* § 141(b). The Secretary must send that tabulation to the President within 9 months of the census date. *Id.* The second is the “[t]abulations of population” for specific areas within the States for the States to use for redistricting. *Id.* § 141(c). The Secretary must send those tabulations to the States within “one year after the decennial census date.” *Id.* Both sets of numbers must be accurate so they can be used for the purposes Congress intended—apportionment and redistricting.

This lawsuit primarily concerns the second set of numbers—the tabulations of population provided to the States for redistricting. Congress created a multi-year process for the Census Bureau and the States to work together to ensure that the Bureau provides the State the population tabulations it needs for redistricting. The process begins “not later than April 1 of the fourth year preceding the decennial census date,” when the Secretary is required to establish criteria for States’ “plan[s] identifying the geographic areas for which specific tabulations of population are desired.” *Id.* “Such criteria shall include requirements which assure that such plan shall be developed in a nonpartisan manner.” *Id.*

Then, “not later than 3 years before the decennial census date,” the “officers or public bodies having initial responsibility for the legislative apportionment or districting of each State may ... submit to the Secretary a plan identifying the geographic areas for which specific tabulations of population are desired.” *Id.* These plans must meet the criteria set by the Secretary. If they do not, the Secretary “shall consult to the extent necessary with such officers or public bodies” to bring the plan into compliance. *Id.* Alabama timely submitted, and the Secretary approved, a plan

identifying the geographic areas for which tabulations of population are needed. *See* Ex. 1, Declaration of Donna Overton Loftin, at 2 (“Loftin Declaration”); Ex. 2, Declaration of Sen. James McClendon, at 1-2 (“McClendon Declaration”).

After plans are finalized and approved, “[t]abulations of populations for the areas identified in any plan approved by the Secretary shall be completed by him as expeditiously as possible after the decennial census date and reported to the Governor of the State involved and to the officers or public bodies having responsibility for legislative apportionment or districting of such State.” 13 U.S.C. § 141(c). The “tabulations of population of each State requesting a tabulation plan, and basic tabulations of population of each other State, shall, in any event, be completed, reported, and transmitted to each respective State within one year after the decennial census date.” *Id.*

The Act defines “decennial census date” as April 1 of the year in which the decennial census takes place. *Id.* § 141(a). One year from April 1 is March 31 of the following year. So, for the 2020 decennial census, the Secretary “shall” transmit the tabulations of populations for redistricting by March 31, 2021.

**B. Alabama Relies on the Census Bureau’s “Tabulations of Population” to be Accurately and Timely Reported.**

Article I of the Constitution grants States the authority to regulate the “Times, Places and Manner of holding Elections for Senators and Representatives.” U.S. Const. art. I, § 4, cl.1. This language confers upon States the “authority to provide a complete code for congressional elections ...; in short, to enact the numerous requirements as to procedure and safeguards which experience shows are necessary in order to enforce the fundamental right involved.” *Smiley v. Holm*, 285 U.S. 355, 366 (1932).

Federal law informs Alabama’s State-law reapportionment and redistricting requirements. Pursuant to the U.S. Constitution, States must draw congressional districts equal in number to the

number of seats the States are apportioned based on their populations. *See* U.S. Const., art. I, § 2, cl.3. Additionally, the one-person-one-vote principle requires that States draw legislative districts that are nearly equivalent in population. *See Evenwel v. Abbott*, 136 S. Ct. 1120, 1123-24 (2016). To abide by these principles, States like Alabama rely on the Census Act’s guarantee that the Bureau will provide timely and accurate redistricting data. *See* 13 U.S.C. § 141(c). Alabama relies on this data for many different functions, including legislative and congressional apportionment. *See* Ex. 1, Loftin Declaration at 2. The reapportionment and redistricting data required under the Census Act thus further Alabama’s sovereign interests in ensuring its representative districts are fairly drawn and that they are sufficiently equal in population to meet the Constitution’s requirements.

Alabama has expressly tied its redistricting processes to the Bureau’s decennial census numbers. *See* Ala. Const. art. IX, §§ 197-200. So have many other States.<sup>3</sup> The Alabama Constitution, for instance, requires that the State Legislature use the number of inhabitants, as reported by the Census Bureau, to apportion the seats in the State House and State Senate. *See* Ala. Const. art. IX, §§ 197-98. The Legislature must also conduct legislative redistricting based on the Census Bureau’s tabulations. *See* Ala. Const. art. IX, §§ 199-200. The Alabama Legislature cannot practically conduct these tasks without accurate redistricting data from the Census Bureau. *See* Ex. 1, Loftin Declaration at 2 (“Because each of Alabama’s electoral districts is based on population as reported by the decennial census results, the [Permanent Legislative Committee and Reapportionment] cannot redistrict until these results are released.”).

---

<sup>3</sup> *See Census Data Snafu Upends 2022 Elections*, Politico (Mar. 1, 2021), <https://perma.cc/DZ5N-275Y> (“At least nine states have constitutional or statutory deadlines to redraw their maps, according to the National Conference of State Legislatures ....”).

Redistricting is not the only election-related process in Alabama tied to the arrival of new census data. As the Deputy Chief of Staff and Director of Elections for the Alabama Secretary of State's Office has attested, once redistricting is completed, "[e]ach of the more than 3 million registered voters in Alabama must be assigned to the correct congressional, State, and local districts." *See* Ex. 3, Declaration of Clay S. Helms, at 2 ("Helms Declaration"). "[O]f course, where a voter lives determines which races the voter can participate in." *Id.* But assigning voters to their correct districts is no small feat. In 50 of Alabama's 67 counties, "the Boards of Registrars perform the reassignment process manually," requiring "officials to pore over maps and lengthy lists of voters to ensure that each voter is correctly assigned to his or her correct precinct." *Id.* at 2. "This task can take a county's Board of Registrars up to 6 months." *Id.* "For example, in 2017, following redistricting litigation, the Alabama Legislature drew remedial House and Senate plans that altered only a portion of the districts in each plan. Even though only some districts were affected, local election officials struggled to complete the district assignment process in in 6 months." *Id.* at 3. By law, though, the voter reassignment process must be complete before the primary election rolls around—and absentee voting begins 55 days before election day. Ala. Code §§ 17-11-5(b); 17-11-12. For Alabama's statewide 2022 primary elections, absentee voting will begin March 30, 2022. *See* Ex. 3, Helms Declaration at 3. If the Bureau were to heed its statutory obligations and deliver the redistricting data no later than March 31, 2021, Alabama's Boards of Registrars shouldn't have a problem reassigning Alabama's registered voters to their correct precincts and districts before absentee voting begins. But the Bureau's delays in delivering the data will force the Legislature to delay redistricting, and the Boards of Registrars will be left with precious little time to assign voters to their new voting districts before voting begins.

The candidates who run in Alabama’s elections also rely on timely and accurate census data. *See* Ex. 3, Helms Declaration at 4. For one, many elected positions in Alabama State government have residency requirements for candidates, *see, e.g.*, Ala. Const. art. IV, § 47, so it is important for these candidates to know where district lines will fall as early as possible. For another, candidates intending to participate in the 2022 primary election may begin soliciting and accepting contributions on May 24, 2021, Ala. Code § 17-5-7(b)(2), and must file a declaration of candidacy by January 28, 2022, *see id.* § 17-13-5(a). And independent candidates and minor political parties must also submit signatures of registered voters who are eligible to vote in the election at issue to achieve ballot access. Ex. 3, Helms Declaration at 4. “The State has faced lengthy litigation in the past when the time for gathering signatures was shortened.” *Id.*

**C. The Census Bureau Adopts a Statistical Method Called “Differential Privacy” That Will Cause the “Tabulations of Population” Used for Redistricting to be Inaccurate.**

*1. The Census Bureau Has Relied on Various Disclosure Avoidance Methods to Successfully Protect the Privacy of Census Respondents in the Past.*

Congress requires that the Census Bureau protect the private information of those who participate in the decennial census. *See* 13 U.S.C. § 9. In particular, the Bureau may not “make any publication whereby the data furnished by any particular establishment or individual ... can be identified.” *Id.* § 9(a)(2). The Bureau has used a number of disclosure avoidance methods to successfully protect the identity of census respondents in recent censuses.<sup>4</sup>

For example, in the 2010 census, first, and most basically, before releasing any files with data at the respondent level (“microdata”), the Bureau removed the direct identifiers of the respondents—their names, addresses, telephone numbers, and the like.<sup>5</sup>

---

<sup>4</sup> *See generally* Laura McKenna, U.S. Census Bureau, *Research & Methodology Directorate: Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Census of Population and Housing Public Use Microdata Samples* (Apr. 2019), <https://perma.cc/9LBN-5BWV>.

<sup>5</sup> *Id.* at 4.

Second, the Bureau used topcoding and bottom-coding to mask outliers in data involving continuous variables, “such as age and dollar amounts.”<sup>6</sup> “When topcoding, the top 0.5 percent of all values or the top 3.0 percent of all nonzero values (whichever effects the least amount of records) are cut off” and replaced with the topcode cut-off value “or the mean or interpolated median of all topcoded values.”<sup>7</sup> So, for example, someone whose age is 95 may have her age instead recorded as 90 to ensure that she does not stick out in an uncrowded census block. Bottom-coding works the same way, just on the other end of the distribution.<sup>8</sup>

Third, the Bureau set minimal weighted-population thresholds that needed to be met before it released data regarding that population. For example, categorical variables needed to have “at least 10,000 people nationwide in each published category.”<sup>9</sup> If the threshold was not met for a certain category, the category would be combined with another one (or ones) until it was. The categories would then be recoded as a broader interval and published that way.<sup>10</sup> The Bureau also recoded or rounded the numbers for certain “categorical and continuous variables,” such as property taxes and responses involving certain dollar amounts.<sup>11</sup>

Fourth, and most significantly, the Bureau used data swapping of household data in the 2000 and 2010 censuses to protect the identity of records with a high risk of disclosure.<sup>12</sup> Data swapping works like this: “Consider a census block with just 20 people in it, including one Filipino American. Without any disclosure avoidance effort, it might be possible to figure out the identity

---

<sup>6</sup> *Id.*

<sup>7</sup> *Id.*

<sup>8</sup> *Id.*

<sup>9</sup> Amy Lauger et al., U.S. Census Bureau, *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research 2* (Sept. 26, 2014), <https://perma.cc/2UXQ-SAFI>

<sup>10</sup> *Id.*

<sup>11</sup> See McKenna, *supra*, at 4.

<sup>12</sup> See Nat’l Conf. of State Legislatures, *Differential Privacy for Census Data Explained* (Feb. 1, 2021), <https://perma.cc/DA93-36GA>.



of that individual. With data swapping, the Filipino American’s data might be swapped with that of an Anglo American from a nearby census block—a census block where other Filipino Americans reside. The details for the person would be aggregated with others, and therefore not identifiable, and yet the total population in both census blocks would remain accurate.”<sup>13</sup>

Data swapping was the “primary way of protecting Census 2010 ... tabular data products.”<sup>14</sup> Notably, not all data are swapped between households when this technique is used. Rather, “[o]nly records which [a]re unique in their block based on a set of key demographic variables” are swapped.<sup>15</sup> All other variables—most importantly, the population numbers—are left undisturbed. Additionally, because the swaps typically occur within the same general geographic area—“for example, across [census] tracts but within the same county”<sup>16</sup>—the error rate (that is, the number of “false” household reports caused by swapping) is reduced as census data are viewed at higher levels of geographic scope. In this way, race data, for instance, can remain relatively “true” for a state or federal legislative district, even if the household records within that district are swapped with those in nearby census blocks. Errors are pushed to the geographic boundaries.

Fifth, the Bureau has also used partially synthetic data to protect records at group quarters for which data swapping is not an option. (Records from a nursing home group quarters, for example, cannot be swapped for those at a nearby college dorm.)<sup>17</sup> To create partially synthetic data, the data are modeled according to a general linearized model and records that may cause a disclosure risk are flagged. “[Th]e variable values that are causing the disclosure risk problem in a given

---

<sup>13</sup> *Id.*

<sup>14</sup> Laura Zayatz et al., U.S. Census Bureau, *Disclosure Avoidance for Census 2010 and American Community Survey Five-year Tabular Data Products* 11 (Nov. 23, 2009), <https://perma.cc/GF4V-QTVA>.

<sup>15</sup> *Id.* at 4.

<sup>16</sup> McKenna, *supra*, at 5

<sup>17</sup> *See id.*

unique record are then blanked and replaced with values generated from the model.”<sup>18</sup> Importantly, “[g]eography and type of [group quarters] are never altered, and the numbers of people of less than age 18 and age 18 or more are never changed.”<sup>19</sup> Thus, States are still given an accurate picture of how many people are present at each address and whether they are of voting age.

These protections worked “extremely well.”<sup>20</sup> “Indeed, there is not a single documented case of anyone outside the Census Bureau revealing the responses of a particular identified person in public use decennial census or [American Community Survey] data.”<sup>21</sup>

2. *The Census Bureau Adopts “Differential Privacy” for the 2020 Census.*

In September 2017, the U.S. Census Bureau announced at its Scientific Advisory Committee meeting that it would be using a disclosure avoidance method called “differential privacy” for the 2020 census.<sup>22</sup> John M. Abowd, the Chief Scientist and Associate Director for Research Methodology at the Census Bureau, publicly announced the decision the following August at the Association of Computing Machinery’s Special Interest Group on Knowledge, Discovery, and Data Mining’s annual conference in London.<sup>23</sup> Three months after that, differential privacy was added to the fourth (and latest) version of the Bureau’s 2020 Census Operational Plan. *See* Ex. 4, U.S. Census Bureau, *2020 Census Operational Plan: A New Design for the 21st Century—Version 4.0* 135, 139-40 (Dec. 2018).

---

<sup>18</sup> *Id.*

<sup>19</sup> *Id.*

<sup>20</sup> Steven Ruggles et al., *Differential Privacy and Census Data: Implications for Social and Economic Research*, 109 AEA Papers and Proceedings 404 (May 2019), <https://perma.cc/GW29-GNAV>.

<sup>21</sup> *Id.*

<sup>22</sup> *See* Simson L. Garfinkel, U.S. Census Bureau, *Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the 2018 End-to-End Test* (Sept. 15, 2017), <https://perma.cc/4J8B-ZEXM>.

<sup>23</sup> *See* John M. Abowd, U.S. Census Bureau, Presentation to the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, *The U.S. Census Bureau Adopts Differential Privacy* (Aug. 23, 2018), <https://perma.cc/USZ6-ZPLC>.

Differential privacy is a “formal privacy” method that “inject[s] a precisely calibrated amount of noise”—intentional error—“into the data to control the privacy risk of any calculation or statistic.”<sup>24</sup> “[T]he goal of differential privacy is to obscure the presence or absence of any individual, or small group of individuals,” from the dataset.<sup>25</sup> The dataset “is said to be differentially private if by looking at the output, one cannot tell whether any individual’s data was included in the original dataset or not.”<sup>26</sup> To accomplish this goal, data are intentionally skewed by a statistical method to reduce the risk of re-identification of the true responses. *See generally* Ex. 5, Expert Report of Dr. Michael Barber (“Barber Expert Report”) (explaining how differential privacy works).

Under differential privacy, the accuracy of the data is viewed as a direct trade-off with privacy. Because differential privacy results from mathematically scrambling the true numbers, “perfect privacy would result in completely useless data,” while “perfect accuracy would result in releasing the data in fully identifiable form.”<sup>27</sup> The chosen blend of accuracy and privacy results in a measure called the “privacy-loss budget” or “Epsilon” ( $\epsilon$ ). Dialing the epsilon up to infinity results in perfect accuracy but theoretically imperfect privacy, whereas setting the epsilon at zero results in perfect privacy but useless data. *See* Ex. 5, Barber Expert Report at 10-11.

The global privacy-loss budget for the 2020 census has not been set. Nor has there been a formal mechanism for outside input or participation—from the political branches or otherwise—

---

<sup>24</sup> Michael Hawes, U.S. Census Bureau, *Title 13, Differential Privacy, and the 2020 Decennial Census* 22 (Nov. 13, 2019), <https://perma.cc/MRQ2-67WG>; *see also* JASON, *Formal Privacy Methods for the 2020 Census* (Apr. 2020), <https://perma.cc/G8ZM-YNN6>.

<sup>25</sup> *See* Cynthia Dwork, *Differential Privacy: A Cryptographic Approach to Private Data Analysis*, in *Privacy, Big Data and the Public Good* 302-03 (Julia Lane et al., eds., 2014)

<sup>26</sup> Harvard University Privacy Tools Project, *Differential Privacy*, <https://perma.cc/T7NJ-N397> (last visited Mar. 2, 2021).

<sup>27</sup> Hawes, *Title 13, Differential Privacy, and the 2020 Decennial Census*, *supra*, at 22 (cleaned up).

in that decision, even though “[t]he proponents of differential privacy ... have always maintained that the setting of [the privacy-loss budget] is a policy question, not a technical one.”<sup>28</sup> The Bureau did not provide notice in the Federal Register of its decision to adopt differential privacy for the 2020 census. Nor did it otherwise seek public comment before the decision was made. The closest it came was a July 2018 solicitation of feedback “to understand how the public uses decennial census data products”—but that solicitation occurred *after* the Bureau had already determined that it would be using the new method for the 2020 census, and it expressly *excluded* feedback regarding the Bureau’s redistricting products. *See* Soliciting Feedback From Users on 2020 Census Data Products, 83 Fed. Reg. 34,111 (July 19, 2018). Nor did that solicitation mention differential privacy or disclosure avoidance methods in any case. *Id.*

Instead, the Census Bureau’s Data Stewardship Executive Policy Committee is expected to set the global privacy-loss budget in early June 2021.<sup>29</sup> But the Bureau *has* already determined the “invariants”—i.e., unaltered numbers—that it will provide the States for redistricting. They will be (and *only* will be): (1) the total population of each State, (2) the total housing units at the census block level, and (3) the number of group quarters facilities by type at the census block level.<sup>30</sup> (By comparison, “[i]n 2010, at the *block level*, total population, voting age population, total housing units, occupancy status, group quarters count and group quarters type were all held invariant.” *See* Ex. 5, Barber Expert Report at 9 (emphasis added).) All other tabulations—such as how many people live in a census block, or how many of those people identify as a certain race—

---

<sup>28</sup> Simson L. Garfinkel, John M. Abowd, & Sarah Powazek, *Issues Encountered Deploying Differential Privacy* 3 (Sept. 6, 2018), <https://perma.cc/7TZQ-AFTD>.

<sup>29</sup> *See* U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), <https://perma.cc/D6VJ-N5Z3>.

<sup>30</sup> *See* U.S. Census Bureau, *2020 Census State Redistricting Data (Public Law 94-171) Summary File 7-3* (Feb. 2021), <https://perma.cc/9HWC-492T>.

will be skewed intentionally. Not only that, but the variant tabulations will be skewed in a way that affects different populations differently. “Rural areas will see a greater variance from the raw data than urban areas.”<sup>31</sup> “Smaller subpopulations, such as specific racial groups, will be affected more than larger racial or ethnic groups.”<sup>32</sup> And “[t]he impact on states will vary, depending on their overall demographics.”<sup>33</sup>

3. *Differential Privacy is Unnecessary and Unproven in the Census Context.*

The Bureau’s purported reason for using differential privacy is a concern in that the rise in computer power, coupled with the proliferation of databases containing individuals’ personal information, creates a risk that someone could use outside data sources along with information from the census tabulations to re-create individual level data.<sup>34</sup>

To explore this risk, the Census Bureau conducted an internal database reconstruction experiment “that sought to identify the age, sex, race, and Hispanic origin for the population of each of the 6.3 million inhabited census blocks in the 2010 census” from the publicly released tabular data.<sup>35</sup> The analysts were purportedly able to use publicly released 2010 census data to reconstruct individual-level microdata with the block, sex, age, race, and ethnicity characteristics for 46% of the population—meaning that the analysts were able to group those characteristics together correctly, but without personal identifying information like a name or address to match.<sup>36</sup> The analysts then purportedly linked the block, sex, and age characteristics they had reconstructed to commercial databases, which provided possible re-identification matches for 45% of the population.<sup>37</sup> The

---

<sup>31</sup> Nat’l Conf. of State Legislatures, *supra*.

<sup>32</sup> *Id.*

<sup>33</sup> *Id.*

<sup>34</sup> See Hawes, *Title 13, Differential Privacy, and the 2020 Decennial Census*, *supra*, at 13.

<sup>35</sup> Ruggles et al., *supra*, at 404; see Hawes, *Title 13, Differential Privacy, and the 2020 Decennial Census*, *supra*, at 17-18.

<sup>36</sup> Hawes, *Title 13, Differential Privacy, and the 2020 Decennial Census*, *supra*, at 18.

<sup>37</sup> *Id.*

name, block, sex, age, race, and ethnicity characteristics from the commercial data—the putative matches—were then compared to the confidential Census data to see if they had in fact been positive re-identifications. A little over a third of this subset were matches—the race and ethnicity for those characteristic sets had been “learned exactly, not statistically.”<sup>38</sup>

Notably, the experiment did not prove that someone *without* the Census’s confidential database—called the Hundred-percent Detail File—could match the characteristics learned from the published tabular datasets with personal identifying information such as names or Social Security numbers from external databases with any degree of reliability or certainty.<sup>39</sup> In other words, no person engaging in reconstruction can know if her “reconstructed” dataset bears any similarity to the true dataset unless she can cross-reference it with the unredacted Hundred-percent Detail File. But no one outside of the Census Bureau can do that—which is also why no one can run the same experiment the Census did, and why details of the experiment have not been published. Census analysts, therefore, concluded that “the risk of re-identification is small.”<sup>40</sup>

Indeed, as experts outside the Census Bureau explained, the test showed that “the system worked as designed: because of the combination of swapping, imputation and editing, reporting error in the census, error in the identified credit agency file, and errors introduced in the microdata reconstruction, there [wa]s sufficient uncertainty in the data to make positive identification by an outsider impossible.”<sup>41</sup> The existing protections “worked extremely well to” prevent an outside

---

<sup>38</sup> John M. Abowd, U.S. Census Bureau, Presentation to the Am. Ass’n for the Advancement of Science, *Staring Down the Database Reconstruction Theorem* (Feb. 16, 2019); <https://perma.cc/P3YV-FXPG>.

<sup>39</sup> Ruggles et al., *supra*, at 405 (“Reconstructing microdata from tabular data does not by itself allow identification of respondents; to determine who the individuals actually are, one would then have to match their characteristics to an external identified database (including, for example, names or Social Security numbers) in a conventional re-identification attack.”).

<sup>40</sup> Abowd, *The U.S. Census Bureau Adopts Differential Privacy*, *supra*, at 15.

<sup>41</sup> Ruggles et al., *supra*, at 405.

adversary from being able to “positively identify which person provided a particular response.”<sup>42</sup> To date, there has not been “a single documented case of anyone outside the Census Bureau revealing the responses of a particular identified person in public use decennial census or [American Community Survey] data.”<sup>43</sup>

Differential privacy is also unproven in the apportionment context. As Census Bureau officials have noted, “[d]ifferential privacy is less than 15 years old, and most existing mechanisms were created for computer science applications, not the needs of official statistical agencies.”<sup>44</sup> “[T]he situation is analogous to the state of Public Key Cryptography in 1989.”<sup>45</sup> As a result, the Bureau has faced numerous challenges as it seeks to impose a still-developing theory of privacy onto the decennial census. For example, the Bureau has “lacked subject matter experts skilled in the theory and practice of differential privacy,” as well as “toolkits for performing differential privacy calculations for verifying the correctness of specific implementations.”<sup>46</sup> Then there have been the practical challenges of translating a new theory into workable data for users. As one Census Bureau advisor has recognized: “It may be confusing to say that a town has a negative, fractional number of individuals with a particular combination of uncommon attributes.”<sup>47</sup>

---

<sup>42</sup> *Id.* at 404.

<sup>43</sup> *Id.*

<sup>44</sup> Garfinkel et al., *Issues Encountered Deploying Differential Privacy*, *supra*, at 3.1.

<sup>45</sup> *Id.* at 3.2.

<sup>46</sup> *Id.*

<sup>47</sup> Michael B. Hawes, U.S. Census Bureau, *Implementing Differential Privacy: Seven Lessons From the 2020 United States Census*, Harvard Data Science Review (Apr. 30, 2020), <https://perma.cc/DB66-9B5R>. The Bureau fixed this problem by imposing a non-negativity constraint on the algorithm, which in turn makes the results even less accurate. *See* Ex. 5, Barber Expert Report at 13-14 (explaining that “[t]he combination of the non-negativity constraint and population invariants consistently leads to bias increasing counts of small subgroups and small geographic units and decreasing counts of larger subgroups and geographic units.” (citation omitted)).

4. *Differential Privacy Will Result in Inaccurate Population Tabulations.*

Because differential privacy intentionally skews all population numbers save for the total population of each State, this scrambling of the data risks rendering it “essentially unusable and unreliable at geographies below the statewide level for redistricting and other purposes.” Ex. 6, Expert Report of Thomas M. Bryan, *Census 2020: Differential Privacy Analysis Alabama Case Study 4* (“Bryan Expert Report”). In October 2019, the Census Bureau released a set of demonstration data for various census stakeholders to review.<sup>48</sup> The Bureau also released additional demonstration data in May, September, and November of 2020.<sup>49</sup> This data applied differential privacy to the 2010 census data for certain States as a means of testing the novel approach to disclosure avoidance. For the demonstration data products, the Census Bureau set a more conservative privacy-loss budget than it expects will be set for the 2020 census—meaning that the demonstration data will have more “noise (error) than should be expected in the final 2020 Census data products.”<sup>50</sup> But the final numbers will still be erroneous—and intentionally so. They will just be less wrong than the demonstration numbers were.

The demonstration data have shown that differential privacy—no matter where the epsilon value is set—inhibits a State’s right to draw fair lines. Simply put, differential privacy forces States to draw districts using false numbers about how many and what type of people reside in a census block, block group, tract, or county. Not only that, but as demographer Thomas Bryan notes in his expert report: Differential privacy “has been in development at the Census Bureau for many years, and we are currently in the time frame we would be preparing for the release of the data under

---

<sup>48</sup> See U.S. Census Bureau, 2010 Demonstration Data Products (rev. Apr. 16, 2020), <https://perma.cc/KK5M-KLRL>.

<sup>49</sup> See U.S. Census Bureau, 2020 Disclosure Avoidance System Updates (Feb. 23, 2021), <https://perma.cc/D6VJ-N5Z3>.

<sup>50</sup> *Id.*



statutory timetables. And the Census Bureau has not yet produced a data product that is even remotely usable by the end user community—including state and local governments for the purpose of redistricting.” Ex. 6, Bryan Expert Report at 5.

Indeed, the level of falsity introduced by differential privacy is unlike past disclosure avoidance methods in significant ways—both in kind and in degree. Most significantly, when data swapping was used to protect small populations, the “total population, voting age population, total housing units, occupancy status, group quarters count and group quarters type were all held invariant” at the census-block level. *See* Ex. 5, Barber Expert Report at 9. In other words, the Bureau provided the States the actual number of people the Bureau counted in each census block. No longer. Under differential privacy, the population numbers themselves are manipulated (save for the statewide level). This is a new kind of error being purposefully introduced into redistricting data. The result is that the States will not know where their residents were counted.

And whereas the errors caused by swapping between adjacent census blocks were largely cancelled out as one looks at higher census geographies<sup>51</sup> because the adjacent blocks are combined together, the same is not true for the errors caused by differential privacy. Those errors can compound as census blocks are combined to form larger census geographies because the population totals and characteristics in adjacent blocks are skewed at random. Unlike in years past, then, “[d]ifferential privacy will mean that, except at the state level, population and voting age population will not be reported as enumerated. And, race and ethnicity data are likely to be farther from

---

<sup>51</sup> Census data are broken up into ever smaller levels of geographic areas called “census geographies.” There are two different classifications of census geography—“on the spine” and “off the spine.” The “on the spine” geographies are, from largest to smallest: Nation, Regions, Divisions, States, Counties, Census Tracts, Block Groups, and Blocks. “Off the spine” geographies are designations for defining other areas of geography for various statistical or other purposes. Some examples of “off the spine” census geographies are: Zip Codes, School Districts, Congressional Districts, Economic Places, Voting Districts, Urban Areas, and Metropolitan Areas.

the ‘as enumerated’ data than in past decades, when data swapping was used to protect small populations.” Ex. 5, Barber Expert Report at 9 (quotation marks and citation omitted).

Examples from the demonstration data prove the point. In the State of Washington, for instance, application of differential privacy “displaced nearly 18% of Washington’s population at the census block level.”<sup>52</sup> When applied to smaller census geographies, the problems became worse. Census blocks with a small number of housing units had much higher populations than were reported by the true 2010 numbers, while blocks with more than 20 housing units had lower populations than they should have had. “In terms of household population, census blocks with only one housing unit had collectively 64,195 more people after applying [differential privacy]. There were also 15,253 people in census blocks that had housing but no population in the original 2010 data. Together, these numbers represent 79,448 people.”<sup>53</sup>

Nor were the falsities introduced by differential privacy evenly distributed across populations. An extreme example is the census block that contains Washington’s Correction Center for Women. In the original 2010 census, the census block was, understandably, approximately 99% female. After the application of differential privacy, the same census block was reported to be only 25% female.<sup>54</sup> Data concerning racial characteristics have been similarly skewed. As the National Redistricting Foundation reported, “initial analyses suggest that the Bureau’s differential privacy proposal can produce inaccurate counts for minority communities by reallocating population from larger minority groups to smaller ones and by geographically dispersing concentrated minority

---

<sup>52</sup> Mike Mohrman, *The Challenge of Differential Privacy: Confidentiality vs. Usability* (Sept. 15, 2020), <https://perma.cc/4FA7-G4EF>.

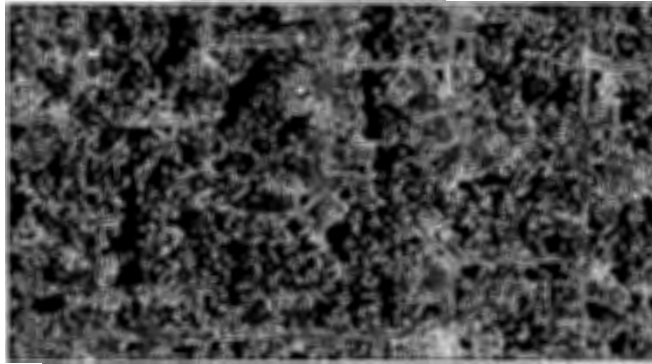
<sup>53</sup> *Id.*

<sup>54</sup> See Mike Mohrman, Letter to Steve Dillingham, Director, U.S. Census Bureau (Feb. 6, 2020), <https://perma.cc/MC3G-62PT>.

populations—precisely the kinds of inaccuracies that would work against the viability of majority-minority districts.”<sup>55</sup>

Such abnormalities appeared in the Alabama data as well. For example, whereas the “2010 Census had 131 children residing in five blocks without adults”—likely reflecting a boarding school or another kind of group quarters for children—the differential privacy algorithm produced over “141,817 children residing in 13,842 blocks without adults.” Ex. 6, Bryan Expert Report at 11. Four of the blocks were reported to have over seventy children residing without adults—though it is clear that two of those blocks consist of single-family neighborhoods:

Block 010730118032035 is a tree lined single family neighborhood on the north side of Birmingham, where it is simply implausible that there are no adults.



---

<sup>55</sup> Nat’l Redistricting Foundation, Letter to Steven Dillingham, Director, U.S. Census Bureau (Apr. 24, 2020), <https://perma.cc/3QK8-65VN>.

Block 010970024001008 is a tree lined single family neighborhood in Mobile south of US 90 where again it is simply implausible that no adults live here.



*Id.* at 13. The application of differential privacy likewise “turned 30,338 blocks with one or more [people of voting age] into blocks with zero [people of voting age.]” *Id.* at 11.

The November 2020 demonstration data also skewed the 2010 tabulations enough to create a population deviation in Alabama’s Congressional districts on a level that courts have found, in other contexts, to violate voters’ equal population rights:<sup>56</sup>

---

<sup>56</sup> See *Veith v. Pennsylvania*, 195 F. Supp. 2d 672 (M.D. Pa. 2002) (three-judge court).

Congressional District	2010 Actual Population	2010 Actual Population Deviation	Differential Privacy Population (Demonstration Data)	Differential Privacy Deviation (Demonstration Data)
1	682820	+1	682747	-73
2	682820	+1	682791	-29
3	682819	-1	682844	+25
4	682819	-1	682820	+1
5	682819	-1	682820	+1
6	682819	-1	682688	-131
7	682820	+1	683026	+206

Ex. 6, Bryan Expert Report at 21.

Then there are the outsized effects differential privacy has on the tabulations of minority populations. For instance, in Alabama there were “19,666 blocks in which [differential privacy] reported zero Hispanic persons of voting age while the 2010 Census reported one or more Hispanic persons of voting age in these same blocks,” and “38,010 blocks in which [differential privacy] reported zero Black Non-Hispanic persons of voting age, while the 2010 Census reported one or more Black non-Hispanic persons in the same blocks.” *Id.* at 12. “Looking in the opposite direction, there were ... 7,384 blocks in which the 2010 Census reported 1 or more Hispanic persons of voting age while [differential privacy] reported zero Hispanic persons of voting age in these same blocks,” and “8,073 blocks in which the 2010 Census reported 1 or more Black non-Hispanic persons of voting age while DP reported zero Black non-Hispanic persons of voting age in these same blocks.” *Id.*

These falsities were reflected in the numbers for Alabama’s 105 state legislative districts. For instance, “[f]or Black / African Americans, there are six districts with both significant numeric and percent differences, which would result in a *significant* change in demographic complexion in these areas” under differential privacy:

<b>Black / African Americans</b>	<b># Error Voting Age Population, Non-voting Age Population</b>	<b>% Error Voting Age Population, Non-voting Age Population</b>
District 25	-376, +199	-5%, +7%
District 35	+414, -73	+8%, -4%
District 62	+421, -451	+5%, -12%
District 64	-262, +440	-3%, 18%
District 68	-93, -533	-1%, -8%
District 70	-361, +287	-2%, +4%

*Id.* at 32.

In sum: If the Census Bureau uses differential privacy, the population tabulations it reports to States for redistricting will be inaccurate.

**D. The Bureau Delays Release of the “Tabulations of Population.”**

In addition to the application of differential privacy, the tabulations of population from the 2020 decennial census will differ in another significant way from past releases. On February 12, 2021, the Census Bureau announced that “it will deliver the Public Law 94-171 redistricting data to all states by Sept. 30, 2021.” Ex. 7, *Press Release: Census Bureau Statement on Redistricting Data Timeline*, U.S. Census Bureau (Feb. 12, 2021), <https://perma.cc/TY9T-UNDM> (the “February 12 Press Release”); *see also* Ex. 8, James Whitehorne, *Census Bureau Statement on Redistricting Data Timeline*, U.S. Census Bureau (Feb. 12, 2021), <https://perma.cc/A2SZ-7L5Q> (the

“Whitehorne Statement,”; and, together with the February 12 Press Release, the “February 12 Decision”). The Bureau acknowledged that the change marked a “delay[] [in] the Census Bureau’s original plan to deliver the redistricting data to the states by March 31, 2021”—the deadline set by Congress in 13 U.S.C. § 141(c). Ex. 7, February 12 Press Release. The Bureau also announced: “Different from previous censuses, the Census Bureau will deliver the data for all states at once, instead of on a flow basis.” *Id.*

**E. Plaintiffs Seek Relief From This Court.**

On March 10, 2021, Plaintiffs filed this suit in the Middle District of Alabama and requested a three-judge panel pursuant to 28 U.S.C. § 2284(a) and Public Law No. 105-119, § 209(b), (d)(1) & (2). *See* Doc. 1. The complaint alleges that Defendants are (1) violating Plaintiffs’ rights under 13 U.S.C. § 141(c) and Public Law No. 105-119, § 209 to accurate tabulations of population for redistricting because differential privacy will cause those tabulations to be inaccurate; (2) creating a substantial risk that Plaintiffs Representative Aderholt, Mr. Green, and Mr. Williams will have their votes in local, state, and federal elections diluted; (3) violating the Administrative Procedure Act (“APA”) because the application of differential privacy is not in accordance with law and contrary to constitutional right, (4) violating the APA because the application of differential is arbitrary, capricious, or constitutes an abuse of discretion, (5) violating Plaintiffs’ rights under 13 U.S.C. § 141(c) by failing to produce the population tabulations by the statutory deadline, (6) violating the APA because Defendants’ delay in producing the population tabulations is contrary to law, and (7) violating the APA because Defendants’ delay in producing the population tabulations is arbitrary and capricious. Plaintiffs also claimed entitlement to a writ of mandamus under 28 U.S.C. § 1361.

### JURISDICTION

The Court has subject-matter jurisdiction pursuant to 28 U.S.C. §§ 1331, 2201(a), and Public Law No. 105-119, § 209(b). Jurisdiction is also proper under the judicial review provisions of APA, 5 U.S.C. § 702.

### LEGAL STANDARD

Plaintiffs are entitled to a preliminary injunction if they show: (1) a likelihood of success on the merits; (2) a likelihood of suffering irreparable harm; (3) that “the threatened injury to the movant outweighs whatever damage the proposed injunction may cause the opposing part[ies]”; and (4) that the injunction would not be adverse to the public interest. *Siegel v. Lepore*, 234 F.3d 1163, 1176 (11th Cir. 2000) (en banc). Alternatively, they are entitled to a writ of mandamus if they have “exhausted all other avenues of relief” and if Defendants owe Plaintiffs “a clear nondiscretionary duty.” *Heckler v. Ringer*, 466 U.S. 602, 616 (1984).

### ARGUMENT

This Court should enter an injunction that (1) enjoins Defendants from applying differential privacy to the tabulations of population Alabama is entitled to for redistricting, and (2) enjoins Defendants from delaying the release of the redistricting data beyond the statutory deadline of March 31, 2021. In the alternative, the Court should grant a writ of mandamus requiring the Secretary to comply with her statutory obligation to provide the State with redistricting data by March 31, 2021.

**I. Plaintiffs Are Entitled To An Injunction Requiring Defendants To Provide Alabama With Timely And Accurate “Tabulations Of Population” Unaffected By The Application Of Differential Privacy.**

Plaintiffs satisfy all four factors needed to obtain a preliminary injunction.

*First*, Plaintiffs are likely to win on the merits. The application of differential privacy violates Congress’s command in 13 U.S.C. § 141(c) that the Secretary provide States “[t]abulations



of population” for redistricting. Rather than provide Alabama tabulations of population for census blocks, towns, counties and the like, she will provide figures generated by the Bureau’s confidential algorithm. And even if those figures could be considered “tabulations of population,” the Census Act requires *accurate* tabulations, not population counts with intentionally added error. Defendants’ use of differential privacy thus violates subsection 141(c) and Plaintiffs’ rights to population tabulations free from manipulation by unlawful statistical methods that affect districting decisions. *See* Pub. L. No. 105-119, § 209(b), (d). This violation will harm the State’s sovereign interest in drawing districts that provide its citizens fair representation and create a substantial risk that the individual Plaintiffs’ votes will be unconstitutionally diluted.

Defendants’ decision to delay the release of the tabulations is likewise unlawful. Congress set the deadline for the Secretary to provide tabulations of population for redistricting by March 31, 2021. Defendants ignored that directive and instead set their own deadline of September 31, 2021. But the deadline set by Congress is not aspirational. It’s the law. Defendants must follow it.

*Second*, Plaintiffs will be irreparably harmed unless the Court enters an injunction. The application of differential privacy to the tabulations of population will violate Plaintiffs’ statutory and constitutional rights, inhibit the State’s right to fairly redistrict, subject the State to the risk of litigation and liability, and likely cost the State federal funding. The Court will be unable to remedy these harms if Defendants deliver population tabulations infected by differential privacy. On the one hand, depending on how differential privacy is applied, the Census Bureau may show that releasing a second set of accurate tabulations would cause significant privacy risks because the two datasets could be compared. In that case, the egg will be impossible to unscramble. On the other hand, if the actual tabulations could one day be released, the prior publication of the false tabulations will mean that States will have to scuttle their redistricting plans drawn from the false

numbers—or face certain litigation for using the second-rate data. Either way, these harms can be avoided if the Court enters an injunction.

*Third*, the benefits of an injunction far outweigh any harm to Defendants. The Bureau plans to deliver the apportionment numbers to the President by April 30, but claims that it then needs five additional months—rather than the three contemplated by statute—to deliver the tabulations of populations to the States. It also has announced that it will be conducting additional testing for differential privacy during this extended time and that the privacy loss budget will not be set until June. This shows that part of the Bureau’s delay is caused by the application of differential privacy in place of tested—and effective—methods of disclosure avoidance. Thus, an injunction that requires Defendants to set aside differential privacy *and* release the population tabulations in a timely manner will minimize harm to Defendants by working together. Reverting to past disclosure avoidance methods will not be difficult or time-consuming; Defendants have done it before with great success. These methods will again enable Defendants to meet their statutory obligations to protect respondents’ privacy *while also providing States with actual population tabulations*—something the current plan does not do. And because applying other methods of disclosure avoidance will be quicker than instituting differential privacy, the injunction will also mean that Defendants will be able to release the population tabulations sooner—more in accord with their statutory obligations.

*Fourth*, an injunction serves the public interest. As a result of Defendants’ actions, States will be forced to redistrict using data that purposefully place people in the wrong place; voters will face a substantial risk that their votes will be diluted; elections will likely be affected; and federal and state governments risk allocating resources to the wrong places. On top of all that, absent an injunction, the correct census data may never be known, and the harms will never be remedied. The public interest strongly favors an injunction.

**A. Plaintiffs Are Likely To Prevail on the Merits.**

*1. The Application of Differential Privacy Violates the Census Act, Individual Plaintiffs' Constitutional Rights, and Plaintiffs' Rights Under Public Law No. 105-119, § 209.*

Defendants' application of differential privacy is unlawful because it violates Congress's command in 13 U.S.C. § 141(c) for the Secretary to provide Alabama accurate tabulations of population to the State for redistricting. It also violates the constitutional rights of the individual Plaintiffs, who face a substantial risk that their votes will be diluted because of the erroneous data.

1. As explained above, Congress created a multiyear process for States desiring specific redistricting data to work with the Secretary before the decennial census to submit "a plan identifying the geographic areas for which the specific tabulations are desired." *Id.* Alabama submitted such a plan, and the Secretary approved it. *See* Ex. 1, Loftin Declaration at 2. Accordingly, Congress directed that the "[t]abulations of population for the areas identified" in Alabama's plan "shall be completed by" the Secretary and given to the State "as expeditiously as possible after the decennial census date ... [but] in any event ... within one year after the decennial census date." 13 U.S.C. § 141(c). Alabama thus has a statutory right to accurate "[t]abulations of population" for geographic areas specific enough to allow for redistricting.

Defendants' refusal to provide this information violates 13 U.S.C. § 141(c) in at least two ways. First, the numbers Defendants will give Alabama are simply not "[t]abulations of population." "The plain-language meaning of 'tabulation of population' is fairly obvious: one counts the number of persons satisfying some required condition(s) and enters that number into a table."<sup>57</sup>

---

<sup>57</sup> JASON, *Formal Privacy Methods for the 2020 Census* 93 (Apr. 2020), <https://perma.cc/G8ZM-YNN6>. Indeed, "tabulate" has long been understood to mean "[t]o put or arrange in a tabular, systematic, or condensed form." *The Random House College Dictionary* 1337 (revised ed. 1975); *see also Seymour v. Barabba*, 559 F.2d 806, 809 (D.C. Cir. 1977) ("Our understanding of a 'tabulation' is a computation to ascertain the total of a column of figures, or perhaps counting the

Yet while Defendants may “count” the number of persons residing in various census blocks throughout Alabama, they will not enter “*that* number” into the tables. Rather, they will enter alternative numbers generated by the Bureau’s confidential algorithm. But Congress did not give the Bureau authority to report estimates or values that merely bear some relation to sub-state population counts. It required that the actual numbers be reported.

Indeed, by declining to apply differential privacy to the State-level population counts the Secretary must report to the President, Defendants appear to recognize that actual population counts for States equate to the “tabulation of total population by States.” 13 U.S.C. § 141(b). The “historical precedent of using the ‘actual Enumeration’ for purposes of apportionment, while eschewing estimates based on sampling or other statistical procedures” forecloses any other interpretation. *U.S. House of Representatives*, 525 U.S. at 340. It follows that the “[t]abulations of population” referenced in subsection 141(c) must also be the actual population counts. Courts, after all, should “not lightly assume that Congress silently attaches different meanings to the same term in the same or related statutes,” much less the same section of the same statute. *Azar v. Allina Health Servs.*, 139 S. Ct. 1804, 1812 (2019). Under either subsection 141(b) or 141(c), actual population counts are required, and close enough isn’t good enough.

Were there any doubt about this point, the canon of constitutional avoidance resolves it. “Where an otherwise acceptable construction of a statute would raise serious constitutional problems, the Court will construe the statute to avoid such problems unless such construction is plainly contrary to the intent of Congress.” *Edward J. DeBartolo Corp. v. Fla. Gulf Coast Bldg. & Const.*

---

names listed in a certain group.”). While the most liberal definition of “tabulate” may include counting rather than simply reformatting the data into tables and lists, this remains a far cry from statistical manipulation deliberately designed to sow error into population numbers. “Tabulate,” Merriam-Webster’s Collegiate Dictionary 1199 (10th ed. 1993) (“2: to count, record, or list systematically.”).

*Trades Council*, 485 U.S. 568, 575 (1988). Defendants’ interpretation of the Census Act raises serious concerns under the Constitution’s Census Clauses. *See* U.S. Const. Art. I, § 2, cl. 3 & amend XIV, § 2. It is “unquestionably doubtful,” for instance, “whether the constitutional requirement of an ‘actual Enumeration,’ Art. I, § 2, cl. 3, [would be] satisfied” if Defendants applied differential privacy to the population numbers it reports under subsection 141(b). *Cf. U.S. House of Representatives*, 525 U.S. at 346 (Scalia, J., concurring) (applying constitutional-doubt canon to conclude that Census Act prohibits use of statistical sampling). Indeed, that may be one reason Defendants chose *not* to apply differential privacy to those numbers. *See Arizona v. Inter Tribal Council of Az., Inc.*, 570 U.S. 1, 17 (2013) (noting that constitutional-doubt canon applies to agency interpretation of statutes). But subsection 141(c) likewise falls within the Constitution’s ambit: The Enumeration Clause serves the “purposes of apportionment,” and that includes intra-state redistricting. *See Dep’t of Commerce*, 525 U.S. at 328-34. Thus, if the Constitution prohibits Defendants from reporting false numbers for the apportionment of representation in the House of Representatives (which it almost certainly does), it also prohibits Defendants from reporting false numbers to the States for redistricting. At the very least, the constitutional question is raised, and that question can be avoided by construing subsection 141(c) in a way that does not put it in potential conflict with the Constitution.

Notably, pointing to disclosure avoidance methods the Census Bureau has used in the past does not help Defendants evade these points. Differential privacy differs in kind, not just degree. JASON, an independent group of scientists and engineers from whom the Census Bureau has sought third-party review, explains this well. “At the time of the 2010 Census, and with the disclosure avoidance procedures adopted at that time, there seemed to be no significant conflict between the statutory requirements” for Defendants to report accurate “[t]abulations of population”

under subsection 141(c) and to protect the privacy of census respondents under section 9.<sup>58</sup> “Swapping, for example, preserves population counts in any geographical area. To the extent that swapped individuals were matched for other characteristics (e.g., voting age), counts of persons with matched characteristics would also be preserved.”<sup>59</sup> In other words, while swapping may change—slightly—the reporting of certain characteristics, the number of people counted in a certain area is still reported accurately in the tabulations. That is important because it is those numbers from which political power is derived and representation is apportioned.

For the 2020 census, though, Defendants have artificially forced the two statutory provisions into conflict. By choosing to report actual and accurate “counts” only for the total population of each State, the total housing units at the census-block level, and the number of group quarter facilities by type at the census-block level, Defendants are refusing to provide Alabama with tabulations of how many of its residents live where. That may comply with Defendants’ privacy obligations (as did past methods), but it violates Defendants’ obligation to report “[t]abulations of population.” There is no reason Defendants cannot comply with both.

Second, even if the numbers Defendants will report constitute “[t]abulations of population,” subsection 141(c) requires accurate, not deliberately inaccurate, numbers to be provided. Any other reading does violence to Congress’s intent, clear from the text of the statute, that Alabama have a right to “specific tabulations of population” for the “geographic areas” identified in the plan it submitted to the Secretary and which the Secretary approved. Defendants’ decision to apply differential privacy will therefore deprive Alabama “of information which it is entitled to receive.” *U.S. House of Representatives v. U.S. Dep’t of Com.*, 11 F. Supp. 2d 76, 85 (D.D.C.

---

<sup>58</sup> JASON, *Formal Privacy Methods for the 2020 Census*, *supra*, at 93.

<sup>59</sup> *Id.*

1998) (three-judge court), *aff'd*, 525 U.S. 316 (1999); *cf. Fed. Election Comm'n v. Akins*, 524 U.S. 11, 24-25 (1998) (recognizing “informational injury”). Surely, for instance, Defendants would agree that assigning every Alabamian to Birmingham would violate Alabama’s right under subsection 141(c) to tabulations of population for specific geographic areas. While Defendants’ final manipulation of the population counts might not be so ambitious, it will be similarly illegal.

By depriving Alabama of the information it is entitled to receive, Defendants will also impede the State’s sovereign interest in drawing fair districts. This is both a separate harm *and* confirmation of the sort of information Congress requires Defendants to provide: population tabulations that can be used for redistricting. *See U.S. House of Representatives*, 525 U.S. at 332-34 (recognizing that unlawful census methods harm States that “use the population numbers generated by the federal decennial census for federal congressional redistricting” or “for their state legislative redistricting”). This means that, for one, the tabulations must at least provide the State the population figures it needs to comply with the Constitution’s one-person, one-vote requirement. The State must be able to “draw congressional districts with populations as close to perfect equality as possible.” *Evenwel v. Abbott*, 136 S. Ct. 1120, 1124 (2016). To do that, the State must know how many people live where, so the tabulations of populations provided by the Secretary must at least provide those figures. Until this census, the Bureau has met that obligation. Under normal circumstances, “the census count represents the ‘best population data available,’ [and] is the only basis for good-faith attempts to achieve population equality.” *Karcher v. Daggett*, 462 U.S. 725, 738 (1983) (quoting *Kirkpatrick v. Preisler*, 394 U.S. 526, 428 (1969)).

For another, the tabulations must allow Alabama to comply with the Voting Rights Act. One of the purposes of Section 2 of the Voting Rights Act is to prevent the State from drawing

districts that “interact[] with social and historical conditions to cause an inequality in the opportunities enjoyed by black and white voters to elect their preferred representatives.” *Thornburg v. Gingles*, 478 U.S. 30, 47 (1986). Ordinarily, a compact and large minority population should be able to elect its candidate of choice. *See id.* at 50-51 (explaining that one of the tests for liability under Section 2 is whether a minority community is “sufficiently large and geographically compact to constitute a majority in a single-member district”). To further its interest in drawing fair districts, then, Alabama needs both actual population data and accurate racial data.

Despite these obligations, Defendants plan to provide the State with inaccurate tabulations—false numbers—except for three broad categories: (1) Alabama’s total population, (2) the total housing units at the census block level, and (3) the number of group quarter facilities by type at the census block level. All other tabulations will be intentionally skewed by differential privacy. That includes the tabulations that normally show how many people live in a census block and how many of those people identify as a certain race—the precise data the State needs to draw fair districts.

As detailed above, and as amply demonstrated in the Bryan Expert Report, *see* Ex. 6, the demonstration data released by the Census Bureau confirm that differential privacy will cause the tabulations of population to be inaccurate. Neighborhoods full of single-family homes were reported to house only children. *See* Ex. 6, Bryan Expert Report at 12-13. Congressional districts drawn from the demonstration data would likely violate one-person, one-vote. *Id.* at 21. And minority populations were misreported to such an extent that voters’ rights under Section 2 of the VRA would likely be violated if the Legislature relied on the demonstration data to draw legislative districts. *Id.* at 22-34.



True, the Bureau has stated that it intends to set a less conservative privacy loss budget for the final tabulations of population than it did for the demonstration products. That should mean that the final tabulations will have less egregious falsities than the demonstration data have had.<sup>60</sup> But by definition, any application of differential privacy will produce erroneous numbers. That's the entire point. It's just that, according to the Census Bureau, the resulting numbers will be less skewed than they are in the demonstration data—though, of course, there will be no way for anyone outside the Census Bureau to ever confirm that. This violates the Census Act's guarantee that Alabama receive accurate tabulations of population and harms the State's sovereign interest in drawing fair districts based on those tabulations.

2. For similar reasons, the individual Plaintiffs face a “substantial risk” that their constitutional rights will be violated by Defendants' application of differential privacy. *See Susan B. Anthony List v. Driehaus*, 573 U.S. 149, 158 (2014). The equal protection component of the Fifth Amendment's due process clause protects the fundamental right to vote. *See Buckley v. Valeo*, 424 U.S. 1, 93 (1976) (“Equal protection analysis in the Fifth Amendment area is the same as that under the Fourteenth Amendment.”). Absent extraordinary justification, one person's vote in a congressional election must be worth as much as another's; and Congressional districts must “be apportioned to achieve population equality ‘as nearly as practicable.’” *Karcher*, 462 U.S. at 730 (quoting *Wesberry v. Sanders*, 376 U.S. 1, 7-8 (1964)). The “as nearly as practicable standard” requires a “good-faith effort to achieve precise mathematical equality.” *Id.* In practice, this requires States to draw congressional districts to mathematic precision of +/- one person. While State legislative districts need not meet the “as nearly as practicable standard,” they must be drawn to be

---

<sup>60</sup> *See* U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), <https://perma.cc/D6VJ-N5Z3>.

within a total population variation of +/- 5% to be presumptively constitutional. *See, e.g., Brown v. Thomson*, 462 U.S. 835, 842 (1983).

The Census Bureau's decision to apply differential privacy—and thus supply false redistricting data to Alabama—creates a substantial risk that Plaintiffs Representative Aderholt, Mr. Green, and Mr. Williams will have their votes in local, state, and federal elections diluted. All three individual Plaintiffs vote regularly, and all three of them live in districts that could be affected by differential privacy. *See* Ex. 9, Declaration of Camaran Williams; Ex. 10, Declaration of William Green; Ex. 11, Declaration of Rep. Robert Aderholt. Defendants are not using a good-faith effort to provide as precise data as possible, and, as a result, Alabama, along with its subordinate governmental units, will be forced to redistrict and reapportion numerous representative districts, including congressional districts, with intentionally flawed data. Defendants are responsible for this vote dilution, which violates Plaintiffs' constitutional right to equal protection.

3. The reason the tabulations of population will be skewed is because of the application of an unlawful statistical method. In Public Law No. 105-119, § 209(a)(7), Congress recognized that “the use of ... statistical adjustment in conjunction with an actual enumeration to carry out the census with respect to any segment of the population poses the risk of an inaccurate, invalid, and unconstitutional census.” And as a plurality of the Supreme Court has explained, while an “actual Enumeration” “may not be the most accurate way of determining population ... it may be the most accurate way of determining population with minimal possibility of partisan manipulation.” *U.S. House of Representatives*, 525 U.S. at 348-49 (Scalia, J., concurring in part). “To give Congress”—or Defendants—the “power ... to select among various estimation techniques having credible (or even incredible) ‘expert’ support is to give the party controlling Congress the power to distort representation in its own favor.” *Id.* At the very least, basing census figures on actual numbers

helps to prevent the *appearance* of improper manipulation. *Cf. Utah v. Evans*, 536 U.S. 452, 471-72 (2002) (noting that imputation is allowed because—among other reasons—it is not “susceptible to manipulation” and “manipulation would seem difficult to arrange”).

As a result, while 13 U.S.C. § 141(c) created a statutory right for each State to receive accurate “[t]abulations of population,” Congress extended that informational right to “[a]ny person aggrieved by the use of any statistical method in violation of the Constitution or any provision of law ... in connection with the ... decennial census[] to determine the population for purposes of ... redistricting Members in Congress.” Pub. L. No. 105-119, § 209(b). By applying differential privacy to skew the tabulations of population, Defendants violate Plaintiffs’ rights under Section 209.

Under Section 209, an “aggrieved person” “includes—(1) any resident of a State whose congressional representation or district could be changed as a result of the use of a statistical method challenged in the civil action; (2) any Representative or Senator in Congress; and (3) either House of Congress.” Pub. L. No. 105-119, § 209(d). Plaintiffs are such “aggrieved person[s].” Congressman Aderholt is a “Representative ... in Congress.” *See* Ex. 11, Aderholt Declaration at 2. Representative Aderholt, Mr. Green, and Mr. Williams are “resident[s] of” Alabama “whose representation or district could be changed as a result of the use of a statistical method.” *See id.*; Ex. 10, Green Declaration at 1-2; Ex. 9, Williams Declaration at 2-3. And because Congress created the guarantee of accurate “[t]abulations of population” for *States* to use in their redistricting process, *see* 13 U.S.C. § 141(c), Alabama is an “aggrieved person,” too. *See* Pub. L. No. 105-119 § 209(a)(8) (noting that Congress created § 209(b)’s cause of action because “it would be impracticable *for the States* to obtain ... meaningful relief after such enumeration has been conducted” (emphasis added)); *cf. Georgia v. Evans*, 316 U.S. 159, 162 (1942) (“Nothing in the [Sherman]

Act, its history, or its policy, could justify so restrictive a construction of the word ‘person’ ... as to exclude a State” where “[s]uch a construction would deny all redress to a State ... merely because it is a State”); *United States v. Schmidt*, 675 F.3d 1164, 1169-70 (8th Cir. 2012) (finding that State agencies were “persons” under the Mandatory Victims Restitution Act); *see also* Antonin Scalia & Bryan A. Garner, *Reading Law: The Interpretation of Legal Texts* 132-33 (2012) (collecting cases for proposition that “[t]he verb *to include* introduces examples, not an exhaustive list”).

Differential privacy is also an unlawful “statistical method.” “[T]he term ‘statistical method’ means an activity related to the design, planning, testing, or implementation of the use of representative sampling, or any other statistical procedure, including statistical adjustment, to add or subtract counts to or from the enumeration of the population as a result of statistical interference.” Pub. L. No. 105-119, § 209(h)(1). It is clear that differential privacy falls into this category. As Professor Michael Barber explained: “At its core, the process of ensuring privacy is a combination of sampling and constrained optimization. Privacy is introduced into the data by introducing random error through sampling from statistical distributions with parameters set to a desired level of variance (privacy) .... Differential privacy is thus an application of statistical processes and methods to adjust the original counts of the Census to protect the privacy of individual[] records.” Ex. 5, Barber Expert Report at 17.

It follows that because differential privacy is a “statistical method” used in violation of 13 U.S.C. § 141(c), and because Plaintiffs are “aggrieved” by that use, Defendants have violated Plaintiffs’ rights under Public Law No. 105-119, § 209(b).

2. *The Application of Differential Privacy Violates the Administrative Procedure Act.*

The APA requires the Court to “hold unlawful and set aside agency action[s]” that are “arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law,” that are “in excess of statutory jurisdiction, authority, or limitations, or short of statutory right,” and that are “contrary to constitution right.” 5 U.S.C. § 706(2)(A), (B), (C). Defendants’ decision to use differential privacy to manipulate the tabulations of population used for redistricting are all those things.

1. To be reviewable under the APA, the challenged decision must constitute “final agency action.” *Id.* § 704. The Supreme Court has created a two-part test to determine whether this is case. “First, the action must mark the consummation of the agency’s decisionmaking process—it must not be of a merely tentative or interlocutory nature. And second, the action must be one by which rights or obligations have been determined, or from which legal consequences will flow.” *U.S. Army Corps of Eng’rs v. Hawkes Co.*, 136 S. Ct. 1807, 1813 (2016) (quoting *Bennett v. Spear*, 520 U.S. 154, 177-78 (1997)). “The core question is whether the agency has completed its decisionmaking process, and whether the result of that process is one that will directly affect the parties.” *Franklin v. Massachusetts*, 505 U.S. 788, 797 (1992).

Both conditions are met here. First, the Census Bureau has completed its decisionmaking process with regard to whether to apply differential privacy to the 2020 decennial census. The Bureau announced that decision in September 2017, and it was added to the 2020 Census Operational Plan in December 2018. *See* Ex. 4, U.S. Census Bureau, *2020 Census Operational Plan: A New Design for the 21st Century—Version 4.0* 135, 139-40 (Dec. 2018). That Plan states: “The disclosure avoidance methodology that *will be implemented* for the 2020 Census is known as differential privacy. Differential privacy is the scientific term for a method that adds “statistical noise”

to data tables we publish in a way that protects each respondent's identity." *Id.* at 139 (emphasis added). The Plan also notes: "[T]he Census Bureau *is implementing* the new differential privacy method. This new methodology will be tested and implemented with the 2018 End-to-End Census Test." *Id.* at 140 (emphasis added). These statements demonstrate that the decision to use differential privacy has been made—done, final. It is not “merely tentative or interlocutory in nature.” *Bennett*, 520 U.S. at 178. To be sure, the Bureau has yet to set the privacy loss budget it will use—that decision is still in the works. But the privacy loss budget—the epsilon—is immaterial. Plaintiffs claim that the application of differential privacy itself—no matter the epsilon—is unlawful. That decision is ripe for review.

Second, it is certain that “legal consequences will flow” from the Census Bureau’s decision to use differential privacy. The decision will cause the Secretary to breach her duty under subsection 141(c) to provide Alabama with accurate tabulations of population for redistricting. It will harm Plaintiffs’ rights under Public Law No. 105-119, § 209(b). And it will force Alabama to draw congressional and legislative districts without accurate data, thus creating a substantial risk that voters like individual Plaintiffs will have their constitutional rights abridged.

The Census Bureau’s decision to adopt differential privacy is contrary to law, contrary to constitutional right, and in excess of statutory authority. *See* 5 U.S.C. § (2)(A), (B), (C). It must be set aside. For the reasons already discussed, the application of differential privacy to the population tabulations given to the States is “inconsistent with the statutory mandate” of 13 U.S.C. § 141(c) and would “frustrate the policy that Congress sought to implement.” *Fed. Election Comm’n v. Democratic Senatorial Campaign Comm.*, 454 U.S. 27, 32 (1981). It would also create a substantial risk that individual Plaintiffs will have their equal protection rights violated. Accordingly, Defendants’ decision violates the APA.

3. The decision violates the APA for another reason: It is arbitrary, capricious, or an abuse of discretion. 5 U.S.C. § 706(2)(A). Agency action is arbitrary and capricious if the agency “has relied on factors which Congress has not intended it to consider, entirely failed to consider an important aspect of the problem, offered an explanation for its decision that runs counter to the evidence before the agency, or is so implausible that it could not be ascribed to a difference in view or the product of agency expertise.” *Motor Vehicle Mfrs. Ass’n v. State Farm Mut. Auto. Ins. Co.*, 463 U.S. 29, 43 (1983).

Here, the Bureau failed to consider and respond to the impact that its decision to adopt differential privacy will have on the States, including Alabama, which rely on the Bureau’s provision of accurate tabulations of population for redistricting and other purposes. But when agencies are not writing on a blank slate—which, given States’ long reliance on accurate redistricting data, the Bureau was not—they must “assess whether there were reliance interests, determine whether they were significant, and weigh any such interests against competing policy concerns.” *Dep’t of Homeland Sec. v. Regents of the Univ. of Cal.*, 140 S. Ct. 1891, 1915 (2020) (citation omitted). Yet the Bureau did not do this. True, the Bureau sought general feedback “to understand how *the public* uses decennial census data products” as it considered whether to “reduce the amount of detailed data” that is released. *See* 83 Fed. Reg. at 34,111 (emphasis added). But the Bureau specifically explained that it was “*not* seeking feedback on apportionment counts and redistricting data products.” *Id.* (emphasis added). The Bureau drastically changed the nature of the redistricting products without soliciting input from the users of those products—the States.

Not only that, but the Bureau’s explanation for why differential privacy is needed runs counter to the evidence and the Secretary’s statutory mandates. For instance, the Bureau claims

that the confidentiality requirements of 13 U.S.C. § 9 dictate the use of differential privacy.<sup>61</sup> But the Bureau has not shown that traditional disclosure avoidance methods like data swapping are insufficient to meet that need. In reality, those methods have worked extremely well. *See* Ex. 6, Bryan Expert Report at 41 (noting that “w[hile the threat of a confidentiality breach is always present, the Census Bureau has not reported any such breaches from prior census data releases”). And importantly, the other methods of disclosure avoidance allow the Secretary to meet her statutory mandate under subsection 141(c) to deliver accurate tabulations of population to the States—which differential privacy will prevent her from doing.

Even assuming that swapping and past disclosure avoidance methods present some level of privacy risk, the Bureau has not explained how the privacy risk under differential privacy will compare. In other words, the Bureau has not shown that differential privacy works better than the disclosure avoidance methods that were used in 2010. But such a determination is needed for the Bureau to make a rational decision about its adoption. And regardless, even if the Bureau did show that differential privacy works better than traditional disclosure avoidance methods, differential privacy cannot eliminate all risks to privacy without making the data completely worthless. Which means that differential privacy and past practices are, at most, simply in different places on the same sliding scale under 13 U.S.C. § 9. Yet the Bureau has also not explained why § 9 would be violated if it chooses the privacy risk associated with past disclosure avoidance methods but will not be violated if it adopts differential privacy. That, too, demonstrates an arbitrary decisionmaking process. What is clear, though, is that past methods do not violate the Secretary’s obligations to report accurate tabulations of population under subsection 141(c), whereas differential privacy will result in such a violation.

---

<sup>61</sup> *See* Abowd, *The U.S. Census Bureau Adopts Differential Privacy*, *supra*, at 9.



Moreover, while the Census Bureau adopted differential privacy because of concerns caused by big data, nothing that the Census Bureau publishes *by itself* can even theoretically lead to the disclosure of confidential information if the Bureau applied the disclosure avoidance methods it has used in the past. Confidentiality is only implicated—in theory—when a recipient of census data uses the information published by the Bureau *together with* other datasets. *See* Ruggles et al., *supra*, at 405 (“Reconstructing microdata from tabular data does not by itself allow identification of respondents; to determine who the individuals actually are, one would then have to match their characteristics to an external identified database (including, for example, names or Social Security numbers) in a conventional re-identification attack.”); *see also* Ex. 5, Bryan Expert Report at 59 (explaining that reidentification can occur “when public data are linked to other external data sources” (citation omitted)). Even then, no person outside the Census Bureau attempting to reconstruct the Census Bureau’s dataset can know if she was successful unless she has access to the confidential Hundred-percent Detail File—which no one outside the Bureau does. In any event, if the Census Bureau is concerned about publishing too much information that can be “linked” to other datasets, it could simply publish less information. Reducing the cross tabulations of data tables or reducing the breakdowns of data tables into fewer cells would both serve this purpose. Reducing the amount of information released in other census datasets would, too. The Bureau has not explained why it chose a method that would harm the States, the people, and violate Congress’s command when many other options are available to it that do no such harms.

In sum, when adopting differential privacy the Bureau did not consider Plaintiffs’ reliance interests, misinterpreted the confidentiality requirements of § 9, failed to explain why past methods of disclosure avoidance are inadequate, and adopted a new statistical method for protecting privacy that would ensure that the Secretary would violate her obligations under subsection 141(c). That

decision was arbitrary and capricious, constitutes an abuse of discretion, and therefore violates the APA.

3. *The February 12 Decision Violates the Census Act.*

Turning now to the Census Bureau's decision to delay releasing the tabulations of population beyond the March 31 deadline, that decision was also unlawful under the Census Act. The violation is simple and clean cut.

First, subsection 141(c) imposes a mandatory deadline on the Secretary to deliver the re-districting tabulations of population to the States by March 31. It states:

Tabulations of population for the areas identified in any [State] plan approved by the Secretary shall be completed by him as expeditiously as possible after the decennial census date and reported to the Governor of the State involved and to the officers or public bodies having responsibility for legislative apportionment or districting of such State, except that such tabulations of population of each State requesting a tabulation plan, and basic tabulations of population of each other State, *shall, in any event, be completed, reported, and transmitted to each respective State within one year after the decennial census date.*

13 U.S.C. § 141(c) (emphasis added). The Act defines “decennial census date” as April 1 of the year in which the decennial census takes place. *Id.* § 141(a). The one-year window from April 1 closes March 31 of the following year. For the 2020 decennial census, then, the Secretary “shall” transmit the tabulations of populations for redistricting by March 31, 2021.

Second, on February 12, 2021, the Census Bureau announced that “it will deliver the Public Law 94-171 redistricting data to all states by Sept. 30, 2021”—not by March 31. Ex. 7, February 12 Press Release at 1.

Ergo, the Secretary will violate subsection 141(c).

The deadline set by Congress is mandatory. According to the statute, the Secretary “shall, in any event” report the tabulations of population to the States by March 31. 13 U.S.C. § 141(c). Congress uses the word “‘shall’ to impose discretionless obligations.” *Lopez v. Davis*, 531 U.S.

230, 241 (2001). Indeed, “[t]he first sign that the statute imposed an obligation is its mandatory language: ‘shall.’ ‘Unlike the word “may,” which implies discretion, the word “shall” usually connotes a requirement.’” *Maine Cmty. Health Options v. United States*, 140 S. Ct. 1308, 1320, (2020) (quoting *Kingdomware Technologies, Inc. v. United States*, 136 S. Ct. 1969, 1977 (2016)).

Congress did not lift the requirement. The Bureau recognized the mandatory nature of the deadline and asked for an extension, but Congress did not provide one. *See Ross v. Nat’l Urban League*, 141 S. Ct. 18, 19 (2020) (Sotomayor, J., dissenting from grant of stay). And interestingly, the extension the Bureau asked for was to deliver redistricting data “to the states no later than July 31, 2021.” *See* U.S. Census Bureau, *U.S. Department of Commerce Secretary Wilbur Ross and U.S. Census Bureau Director Steven Dillingham Statement on 2020 Census Operational Adjustments Due to COVID-19* (Apr. 13, 2020), <https://perma.cc/C2RG-UXBX>. When Congress failed to act on that request, the Bureau purported to grant the extension itself—plus an extra two months.

There is no opt-out provision. While the Bureau claims generally that “COVID-19 delayed census operations significantly,” Ex. 8, Whitehorne Statement, the reason for violating the statute is legally irrelevant. Congress required the Secretary to complete, report, and transmit the State-redistricting numbers within one year “*in any event.*” 13 U.S.C. § 141(c) (emphasis added). There is no reason to think Congress did not mean what it wrote, or that it was unaware that there could be difficulties down the line. *See, e.g., Conn. Nat’l Bank v. Germain*, 503 U.S. 249, 253-54 (1992) (“We have stated time and again that courts must presume that a legislature says in a statute what it means and means in a statute what it says there.”). COVID-19 cannot excuse the Bureau’s failure to comply with the plain text of the Act.

4. *The February 12 Decision Violates the Administrative Procedure Act.*

The February 12 Decision also violates the APA because it is “not in accordance with law” and is “in excess of statutory jurisdiction, authority, or limitations, or short of statutory right.” 5 U.S.C. § 706(2)(A), (B). It is also arbitrary and capricious. *Id.* § 706(2)(A).

The February 12 Decision constitutes final agency action. The question before the Bureau was whether to comply with the statutory deadline of March 31. The Bureau’s determined that it would not. By its own terms, the Decision memorializes a “final[] ... schedule for the redistricting data,” by which the Bureau “will deliver the redistricting data to all states by Sept. 30, 2021.” Ex. 8, Whitehorne Statement. There is nothing left for the Bureau to decide as to whether it will meet the March 31 deadline. The decision “mark[ing] the consummation of the agency’s decisionmaking progress” is that the Bureau will ignore Congress’s command. *Hawkes Co.*, 136 S. Ct. at 1813 (quoting *Bennett*, 520 U.S. at 177-78).

The “legal consequences” of that decision “will flow” directly to Alabama. *Id.* (quoting *Bennett*, 520 U.S. at 178). The decision guarantees that the Secretary will violate the State’s right under subsection 141(c) to receive the redistricting data it is entitled to by March 31. That, in turn, will affect the State’s redistricting plan and cause all sorts of problems for the 2022 election-cycle—as explained below when discussing the irreparable harm Plaintiffs will suffer absent an injunction.

For the same reasons the February 12 Decision violates the Census Act, it also violates the APA. The Decision is “not in accordance with law” and is “in excess of statutory jurisdiction, authority, or limitations, or short of statutory right,” 5 U.S.C. § 706(2)(A), (C), because it violates the March 31, 2021 deadline set by Congress. The Secretary may not unilaterally amend—let alone defy—federal law. *See* U.S. Const., art. I, § 7; *Merck Sharp & Dohme Corp. v. Albrecht*, 139 S. Ct. 1668, 1679 (2019) (“[F]or an agency literally has no power to act, let alone pre-empt the validly

enacted legislation of a sovereign State, unless and until Congress confers power upon it.” (quoting *New York v. FERC*, 535 U. S. 1, 18 (2002)).

The February 12 Decision is arbitrary and capricious as well. The Bureau knows that States rely on accurate, timely census data for redistricting. *See* Ex. 8, Whitehorne Statement (recognizing that “[s]ome states have statutory or even state constitutional deadlines and processes that they will have to address due to this delay”). So the Bureau could have considered those reliance interested and attempted to deliver apportionment and redistricting numbers to different States “on a flow basis”—as it has in the past—prioritizing the States whose laws rely on timely receipt of census data. Instead, the Bureau adopted an all-at-once approach without explaining why it was departing from past practice. *Id.* This evinces a disregard of the significant reliance interests States have in the timely production of redistricting data, as well as a lack of a well-thought-out response to the problems created by the Bureau’s own delay. *See Dep’t of Homeland Sec.*, 140 S. Ct. at 1915.

Not only that, but the Bureau “offered an explanation for its decision that runs counter to the evidence before the agency, or is so implausible that it could not be ascribed to a difference in view or the product of agency expertise.” *Motor Vehicles Mfrs. Ass’n*, 463 U.S. at 43. Of course, it is understandable that the COVID-19 pandemic caused some of the delay experienced by the Census Bureau this past year. But it does not explain why the Bureau has given itself until the end of September to report the tabulations of population to the States.

The Bureau initially set July 31, 2020, as its deadline for concluding the counting portion of the 2020 census (including its non-response follow up operations); and then, due to the pandemic, pushed the deadline back to September 30, 2020.<sup>62</sup> With both deadlines, the Bureau planned to report apportionment data by December 31, 2020. The Bureau concluded its count on October 15, 2020, two-and-a-half months past its original deadline and 15 days past its adjusted deadline. *See 2020 Census Response Rate Update: 99.98% Complete Nationwide*, U.S. CENSUS BUREAU (Oct. 19, 2020), <https://perma.cc/MFE3-8PDP>. The two-and-a-half month delay past the original July 31 deadline cannot possibly justify a four-month extension for apportionment numbers and a six-month extension for redistricting numbers.

Indeed, the Bureau's unprecedented and unexplained delay is all the less justifiable considering the 2020 census's remarkable success and improvement over its 2010 predecessor. In the Bureau's words: "In all states, the District of Columbia and the Commonwealth of Puerto Rico, more than 99% of all addresses have been accounted for, and in all but one state that number tops 99.9%.... [W]e had not expected to exceed the 2010 self-response rate. ... The Census Bureau was able to meet and overcome many challenges because of our innovative design and use of new technology."<sup>63</sup>

Furthermore, the Census Bureau recently stated it would finalize the Decennial Response File 2 ("DRF2") numbers over the last weekend of February. *See Joint Case Management Statement, Nat'l Urban League v. Coggins*, No. 5:20-CV-05799-LHK (N.D. Cal. Feb. 24, 2021), ECF

---

<sup>62</sup> Albert E. Fontenot, 2020 Census Update, Presentation to the Census Scientific Advisory Committee March 18, 2021 at 2, <https://perma.cc/A4UM-FHCU>.

<sup>63</sup> *See U.S. Census Bureau, 2020 Census Response Rate Update: 99.98% Complete Nationwide* (Oct. 19, 2020), <https://perma.cc/MFE3-8PDP>. *see also* U.S. Census Bureau, *Adapting Field Operations To Meet Unprecedented Challenges* (Mar. 1, 2021), <https://perma.cc/AU4S-9GXC> ("As a result of all these extraordinary efforts, we were able to account for over 99.9% of U.S. addresses in the 2020 Census.").

No. 469, at 3. This means that the Census Bureau has only the Census Unedited File (“CUF”) to complete before reviewing and delivering the final apportionment numbers to the President.<sup>64</sup> That process typically takes about a month.<sup>65</sup> And the final counting prior to presentment generally runs on an even shorter timeline.<sup>66</sup> Thus, waiting almost seven additional months for redistricting numbers implies unusually long CUF and final-review processes, which the Bureau has failed to explain.

In the same vein, the Census Act codifies the expectation that the Bureau can (and will) produce redistricting data from apportionment data within a three-month timeframe. *See* 13 U.S.C. §§ 141(b) (setting nine-month deadline from census date for “tabulation of total population”); 141(c) (setting one-year deadline from census date to “complete[], report[], and transmit[]” tabulations of population for redistricting “to each respective State”). Yet the Bureau granted itself *five* months with which to produce redistricting data following the long-delayed delivery of the apportionment data. *See* Ex. 7, February 12 Press Release (stating that the state-population count will be delivered by April 30 and the redistricting data will be delivered by September 30). And again, the Bureau failed to explain this glaring discrepancy.

It appears that one explanation is the Bureau’s difficulty in implementing differential privacy. Given the existing timelines for implementing differential privacy—the next set of demonstration data will be released by April 30, 2021, and the privacy loss budget is to be set this June—it is likely that the application of differential privacy is contributing to the delay.<sup>67</sup> But the Bureau

---

<sup>64</sup> *See* U.S. Census Bureau, *Census Data Processing 101* (Feb. 11, 2020), <https://perma.cc/E8JK-4S9V>.

<sup>65</sup> *See* Letter from JASON to U.S. Census Bureau at 5, fig. 1 (Feb. 8, 2021), <https://perma.cc/D3RF-TEBA>.

<sup>66</sup> *Id.*

<sup>67</sup> *See* U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), <https://perma.cc/D6VJ-N5Z3>.

may not justify its unlawful delay by citing difficulties with differential privacy. For starters, the Bureau decided to implement differential privacy on its own accord. The Bureau may not, therefore, point to delays resulting from its own initiatives as a legal justification for deliberately ignoring a statutory deadline. Then there's the problem that differential privacy is illegal when applied for redistricting purposes, thus doubly dooming its validity as an excuse for the Bureau's delay in producing redistricting data. The Bureau's unexplained, unreasoned, and unlawful decision to delay the release of the tabulations of population until September 30 is arbitrary and capricious and should be set aside.

**B. Without an Injunction, Plaintiffs Will Be Irreparably Harmed.**

In addition to being unlawful, Defendants' decisions to implement differential privacy and to delay releasing the tabulations of population will irreparably harm Plaintiffs.

*1. The Inaccurate Population Tabulations Will Irreparably Harm Plaintiffs.*

Defendants' application of differential privacy will violate Plaintiffs' statutory and constitutional rights, make lawful redistricting difficult, subject the State to the risk of litigation and liability, and likely cost communities in Alabama federal funding and affect the allocation of State educational funding. These harms and others will follow swiftly on the heels of the Bureau's release of skewed data for at least three reasons.

First, the State needs to begin redistricting promptly and thus will need to make use of the Bureau's second-rate data upon its release. There won't be time to wait to see how this Court or the Supreme Court resolves this case.

Second, if Plaintiffs prevail after the skewed data is released and receive accurate data, they will face at least one of two harms. They will either need to redistrict again with the best available data or face certain litigation over the lines they already drew.



And, third, depending on how the Bureau implements differential privacy, there is a risk that once the skewed population tabulations are delivered, the Bureau will be unable to release the unskewed tabulations without causing serious privacy problems from releasing two datasets that could be compared with each other. Defendants might then claim that Plaintiffs' harms are no longer redressable—and they could be right.

Turning to the harms themselves, the Alabama Legislature has relied on the Census Bureau for decades to provide accurate information that can be used for redistricting. But under differential privacy, the Alabama Legislature will not know the actual number of people, or accurate demographic makeup, in any census geography below the level of the State as a whole. As explained above, that will make it difficult for the Legislature to draw legislative and congressional districts with near-equal populations, as the Constitution requires. It will also impede the State's interest in drawing legislative and congressional districts that protect minority voting rights. The application of differential privacy, for instance, will obscure whether minority populations are packed into districts where their voting strength is diluted or spread across districts where they may not be able to elect the candidate of their choice.

These difficulties make litigation against the State especially likely. *See* Jeff Zalesin, *Beyond the Adjustment Wars: Dealing With Uncertainty and Bias in Redistricting Data*, 130 *Yale L.J. Forum* 186, 187-89 (2020) (noting that “the Bureau’s adoption of a new system for protecting respondents’ privacy by algorithmically adding error to published data” is one reason “the 2020 Census [is] at risk of being the least accurate census in living memory,” and urging courts to “strike down maps as unconstitutionally malapportioned” even when “the Census Bureau’s official data products in isolation would point to the opposite result”). Liability under the Voting Rights Act is especially worrisome because the legal and factual tests for a finding of liability turn, in part, on

past findings of liability. If Alabama is held liable because it was forced by the Census Bureau to use data tainted by differential privacy, it will be even *more* likely in future suits to be found liable under Section 2. Findings of liability under the Voting Rights Act can also potentially subject Alabama, and its subordinate governmental units, to the “bail-in” provisions of Section 3(c), which would subject the relevant jurisdiction to continual judicial monitoring similar to the pre-clearance provisions of Section 5.

Then there is the financial harm of Defendants’ actions. Alabama communities stand to lose federal funding if the population tabulations are inaccurate because numerous federal programs rely upon the population figures<sup>68</sup> collected and reported in the decennial census to distribute funds to state and local governments. In Fiscal Year 2017, for example, 176 federal programs relied on local-level census-derived data to distribute federal funding. Roughly a hundred programs relied on state-level census-derived data. And 39 programs relied on both state and local-level census-derived data.<sup>68</sup> “Forty percent of the[se] programs use[d] census-derived data to determine the geographic areas and households eligible to receive the program’s funding.”<sup>69</sup>

Census-derived eligibility or allocation criteria used by federal programs to distribute funding include an area’s population density (such as rural or urban designation) and its population size (above or below a specified level); the number of persons in rural areas and persons in overcrowded housing; and the category of the geographic area—whether it is large metro, metro, micro, rural, or isolated county.<sup>70</sup> And the two primary determinants of how census-guided federal spending is

---

<sup>68</sup> Andrew Reamer, *Counting for Dollars 2020: The Role of the Decennial Census in the Geographic Distribution of Federal Funds*, Brief 7: Comprehensive Accounting of Census-Guided Federal Spending: Part A: Nationwide Analysis (FY2017), <https://perma.cc/WQT9-DBYQ>.

<sup>69</sup> *Id.*

<sup>70</sup> Reamer, Brief 7: Comprehensive Accounting of Census-Guided Federal Spending: Part A, *supra*.

allocated among States are (1) poverty rate and (2) percentage of the population living in a rural area.<sup>71</sup> Differential privacy will affect the application of every one of these factors. As Representative Aderholt attests, “should differential privacy be implemented, a large number of communities will receive a larger portion of federal funding than intended and the reciprocal number of communities will receive a smaller portion of federal funding than intended.” Ex. 11, Aderholt Declaration at 4. “Differential privacy will therefore make any funding by act of Congress that ties funding to population at the sub-state level unreliable and suspect.” *Id.*

In fiscal year 2017, Alabama received approximately \$13 billion through 55 federal spending programs guided by data derived from the 2010 census.<sup>72</sup> This included approximately \$12 billion in federal financial assistance programs such as Medicaid, student loans, Supplemental Nutrition Assistant Program benefits, and Medicare Part B; \$171 million in federal tax expenditures such as the low income housing tax credit and the new markets tax credit; and \$250 million in federal procurement programs.<sup>73</sup> Yet these expenditures are likely to go to the wrong place in the future because “differential privacy is not applied equally across the entire population.” *See* Ex. 5, Barber Expert Report at 13. Rather, “[p]laces with fewer people (rural locations) and areas with smaller, distinctive populations (minority communities) are more likely to be impacted since these are the places where identification is more concerning, and the application of statistical noise is more likely to have a larger impact on the summary statistics derived from altered data.” *Id.* at 13-14; *see also id.* at 14 (“Infusing noise in the data, in comparison to the current disclosure avoidance system, will produce inaccurate patterns of demographic change with higher levels of error found

---

<sup>71</sup> *See* Andrew Reamer, *Counting for Dollars 2020*, Brief 7: Comprehensive Accounting of Census-Guided Federal Spending: Part B: State Estimates (FY2017), <https://perma.cc/8PWU-TM57>.

<sup>72</sup> Reamer, *Counting for Dollars 2020: Alabama, supra*.

<sup>73</sup> *Id.*

in the calculations for non-Hispanic blacks and Hispanics.” (citation omitted)). Thus, “[t]he Census Bureau’s use of differential privacy will result in an inappropriate distribution of funds because the population totals used to assign those funds will be purposely inaccurate.” Ex. 11, Aderholt Declaration at 4.

The differences caused by differential privacy are especially easy to see in the education context. In Fiscal Year 2016, Alabama received approximately \$341 million from four different federal programs that used census-related data to allocate funding for young children. *See* Ex. 6 Bryan Expert Report at 14. Yet “on average across the unified school district of Alabama, there was nearly a 10 percent error in the number of young children ages 0 to 4,” and a “mean absolute percent error for ages 5 to 17 [of] 2.8 percent.” *Id.* at 16. Importantly, however, these averages affected different school districts differently. For instance:

[T]he 2010 Census reported that Clarke County School District had 1,295 children ages 0 to 4, but after [differential privacy] was applied, the number of children ages 0 to 4 was decreased to only 885. This is a reduction of 410 children, or 32 percent.

According to the National Center for Education Statistics, the average class size for public schools in Alabama is about 20 students. The error of 410 students for Clarke County School Districts amounts to about twenty classrooms. If 410 unexpected students show up in the Clarke County School Districts, that will lead to crowded classrooms. On the other hand, building and staffing 20 classrooms that are unneeded because of inaccurate census data would be problematic. That is why getting accurate data on the school-age population is so important.

*Id.* at 16-17.

So, too, for school-aged populations. The 2010 census reported that 9,548 children ages 5 to 17 lived in the Madison City School District. After differential privacy was applied, “the figure was changed to 8,774.” *Id.* at 17. “This is a decrease of 776, or 9.0 percent.” *Id.* Based on the demonstration data, it is clear that “the level of error introduced [by differential privacy] will result

in a high level of errors for many unified school districts in Alabama for both the pre-school population (ages 0 to 4) and the school-age population (ages 5 to 17).” *Id.* These discrepancies will cause federal dollars to be spent in areas where they are less needed and withheld from the areas that need them most.

These examples are easy to find because accurate census numbers affect so much. Yet while financial harms can usually be remedied (though the other harms suffered by Plaintiffs cannot), these will not be if the Bureau cannot deliver actual tabulations after it releases the skewed data without causing significant privacy concerns. In that case, “we will *never* be able to assess the relative accuracy of the [differential privacy] system used for the 20[20] census by comparing it to the results of a headcount,” *U.S. House of Representatives*, 525 U.S. at 349 (Scalia, J., concurring), for the results of the headcount will never be released. *Cf.* Pub. L. No. 105-119, § 209(a)(8) (recognizing that it is often “impracticable for the States to obtain, and the courts of the United States to provide, meaningful relief” after the census process is complete). And again, if it turns out at the end of this litigation that both tabulations can be released, the State will then be forced to scrap the maps it drew based on the faulty data and begin redistricting again—or face lawsuits for relying on bad data. Either way, Plaintiffs will be irreparably harmed by the application of differential privacy unless this Court enters an injunction.

2. *The Delayed Population Tabulations Will Irreparably Harm Plaintiffs.*

Defendants’ delay in producing the population tabulations will also irreparably harm Plaintiffs. When the federal government prevents a State from applying state law, the State suffers an irreparable harm. *See Maryland v. King*, 133 S. Ct. 1, 3 (2012) (Roberts, C.J., in chambers). The Census Bureau’s February 12 Decision hamstring Alabama’s ability to meet its constitutional obligations and to run its 2022 statewide elections effectively or in accordance with State law. Therefore, it irreparably harms the State.

As explained above, delivering redistricting data on September 30 will also likely leave Alabama's Boards of Registrars at most only four months for reassigning their respective counties' registered voters to their correct precincts and districts. But four months will likely not be enough. The reassignments typically take up to six months because most counties perform the reassignment process manually. *See* Ex. 3, Helms Declaration at 2-3. Requiring the Boards of Registrars to complete the reassignment process on such an abbreviated schedule will result in some or all of the following: "(1) thousands of dollars in unexpected costs incurred by the Boards of Registrars to contract with an entity to assist them in the process; (2) a rushed reassignment process, potentially increasing the likelihood of mistaken reassignments; and (3) less time to notify voters about changes, potentially increasing the likelihood of voter, political party, and candidate confusion." *Id.* at 3-4.

Finally, the Bureau's delay harms candidates like Representative Aderholt by effectively reducing by at least four months the amount of time they can spend campaigning and fundraising. *See* Ex. 3, Helms Declaration at 4-5; Ex. 11, Aderholt Declaration at 5; Ala. Code § 17-5-7(b)(2). As Representative Aderholt puts it: "The Census Bureau's delays have a cascading effect on my bid for reelection. The problem is all the more acute in Alabama's case as, based on estimates, Alabama may lose a congressional district[,] ... which w[ould] result in a myriad of additional complications when the new districts are redrawn. However, in any event, the census delays will result in less time for me to educate voters as to my policy positions, campaign amongst the voters, and introduce myself to any new voters." Ex. 11, Aderholt Declaration at 5.

**C. The Benefits of an Injunction Far Outweigh the Costs.**

The next factor is whether "the threatened injury to the movant outweighs whatever damage the proposed injunction may cause the opposing part[ies]." *Siegel*, 234 F.3d 1163 at 1176. It does.

Plaintiffs request an injunction that does two things: enjoin Defendants from applying differential privacy to skew the tabulations of population given to the States, and enjoin Defendants from delaying the release of those tabulations beyond the statutory deadline. Because the application of differential privacy is likely contributing to the delay, this relief works in tandem to allow Defendants to meet their twin obligations under subsection 141(c) to provide the States with accurate *and* timely tabulations of population.

To be sure, such relief will cause Defendants to change course. But that Defendants will be forced to stop violating the law is hardly reason to avoid issuing an injunction. And in any event, the requested relief is reasonable. The Bureau has other methods of disclosure avoidance at its disposal. Applying them here will not be overly costly or time-consuming. In fact, it is likely to be far quicker than implementing differential privacy would be. “The Census Bureau has this methodology ‘on the shelf’ and should have immediate access to sufficient human capital in the form of staff and contract experience required to use it in a short period of time.” Ex. 6, Bryan Expert Report at 41.

**D. An Injunction Will Serve the Public Interest.**

Finally, an injunction will serve the public interest. Federal law requires the Secretary to provide both accurate and timely tabulations of population to the States to use for redistricting. The Secretary is shirking both responsibilities. The effect of that dereliction is substantial and widespread. Voters face a substantial risk that their votes will be diluted as States are forced to rely on false numbers to redistrict; States themselves are deprived of tabulations they are entitled to; elections will likely be upended; and federal and state governments risk allocating resources to the wrong places.

Underlying all those harms is this: Absent an injunction, States—already short on time because of Defendants’ delay—will begin redistricting using the faulty numbers as soon as they

come out. If Plaintiffs ultimately prevail in this litigation, the actual tabulations may eventually be released. If they are, States will be forced to throw away the maps they just drew and start again using the newly available “best population data.” *Karcher*, 462 U.S. at 738 (citation omitted). If they are not, perhaps because releasing the accurate tabulations along with their skewed versions will present real privacy risks, then all the States will be left with is the Bureau’s word that the deviations in the final tabulations were not as bad as they were in the demonstration data. There will be no way to confirm that, of course, and no way to know that the numbers were not improperly manipulated—or will not be improperly manipulated in the future. *Cf. U.S. House of Representatives*, 525 U.S. at 348-49 (Scalia, J., concurring in part) (warning that the application of unlawful statistical methods to census data carries with it the “possibility of partisan manipulation” and the “power to distort representation”). Regardless, then, an injunction should issue to prevent either form of harm.

## **II. In The Alternative, The Court Should Issue A Writ Of Mandamus.**

If the Court determines that it cannot provide the needed relief through an injunction, it should provide partial relief through a writ of mandamus requiring the Secretary to meet the statutory deadline of March 31 to deliver the tabulations of populations for redistricting to the States.

“The district courts shall have original jurisdiction of any action in the nature of mandamus to compel an officer or employee of the United States or any agency thereof to perform a duty owed to the plaintiff.” 28 U.S.C. § 1361. A court may grant a writ of mandamus to a plaintiff who has “exhausted all other avenues of relief” for enforcing “a clear nondiscretionary duty” that the defendant owes to it, *Heckler v. Ringer*, 466 U.S. 602, 616 (1984), and if issuance of the writ is “appropriate under the circumstances,” *Cheney v. U.S. Dist. Ct.*, 542 U.S. 367, 381 (2004).

Here, if the Court determines the State is not entitled to an injunction, then the State has no other avenue of relief to keep the Secretary from breaching a clear nondiscretionary duty. The



Secretary's failure to act in a timely fashion will cause the irreparable harms discussed above, and issuance of "the writ is appropriate under the circumstances." *Cheney*, 542 U.S. at 381. Therefore, if the Court determines that the State is unable to obtain injunctive relief, the Court should issue a writ of mandamus requiring the Secretary to comply with the March 31 deadline imposed by Congress.

**CONCLUSION**

The Court should grant injunctive relief or, in the alternative, issue a writ of mandamus.

Dated: March 11, 2021

Respectfully submitted,

STEVE MARSHALL  
*Attorney General of Alabama*

*/s/ Edmund G. LaCour Jr.*

Edmund G. LaCour Jr. (ASB-9182-U81L)  
*Solicitor General*

A. Barrett Bowdre (ASB-2087-K29V)  
*Deputy Solicitor General*

James W. Davis (ASB-4063-I58J)  
Winfield J. Sinclair (ASB-1750-S81W)  
Brenton M. Smith (ASB-1656-X27Q)  
*Assistant Attorneys General*

STATE OF ALABAMA  
OFFICE OF THE ATTORNEY GENERAL  
501 Washington Ave.  
Montgomery, AL 36130  
Telephone: (334) 242-7300  
Fax: (334) 353-8400  
Edmund.LaCour@AlabamaAG.gov  
Barrett.Bowdre@AlabamaAG.gov  
Jim.Davis@AlabamaAG.gov  
Winfield.Sinclair@AlabamaAG.gov  
Brenton.Smith@AlabamaAG.gov

*Counsel for the State of Alabama*

*/s/ Jason Torchinsky (with permission)*

Jason B. Torchinsky\*  
Jonathan P. Lienhard\*  
Shawn T. Sheehy\*  
Phillip M. Gordon\*

HOLTZMAN VOGEL JOSEFIAK  
TORCHINSKY, PLLC  
15405 John Marshall Hwy  
Haymarket, VA 20169  
(540) 341-8808 (Phone)  
(540) 341-8809 (Fax)  
Jtorchinsky@hvjt.law  
Jlienhard@hvjt.law  
Ssheehy@hvjt.law  
Pgordon@hvjt.law

*\*pro hac vice application to be filed*

*Counsel for Plaintiffs*

**CERTIFICATE OF SERVICE**

I hereby certify that on March 11, 2021, I hand-filed the foregoing with the Clerk of the Court. I further certify that I have on this date mailed a copy of the foregoing to the following parties:

U.S. Department of Commerce  
1401 Constitution Ave. NW  
Washington, DC 20230

Secretary Gina M. Raimondo  
Secretary of Commerce  
U.S. Department of Commerce  
1401 Constitution Ave. NW  
Washington, DC 20230

U.S. Census Bureau  
4600 Silver Hill Road  
Washington, DC 20233

Ron S. Jarmin  
Acting Director  
U.S. Census Bureau  
4600 Silver Hill Road  
Washington, DC 20233

Merrick Garland  
Attorney General  
U.S. Department of Justice  
950 Pennsylvania Ave. NW  
Washington, DC 20530-0001

Sandra J. Stewart  
Acting U.S. Attorney  
United States Attorney's Office for the Middle District of Alabama  
131 Clayton Street  
Montgomery, AL 36104

I further certify that I served a copy of this motion by e-mail upon:

Brad P. Rosenberg  
Assistant Branch Director  
United States Department of Justice  
Civil Division, Federal Programs Branch  
Brad.Rosenberg@usdoj.gov

James DuBois  
Assistant United States Attorney  
Civil Chief  
United States Attorney's Office for the Middle District of Alabama  
james.dubois2@usdoj.gov

  
/s Edmund G. LaCour Jr.  
Counsel for State of Alabama

RECEIVED

2021 MAR 11 P 4:49

DEBRA P. HACKETT, CLP  
U.S. DISTRICT COURT  
MIDDLE DISTRICT ALA

## Exhibit 5

**UNITED STATES DISTRICT COURT FOR THE  
MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION**

THE STATE OF ALABAMA; ROBERT  
ADERHOLT, Representative for Alabama's  
4th Congressional District, in his official and  
individual capacities; WILLIAM GREEN;  
AND CAMARAN WILLIAMS,

Plaintiffs,

v.

UNITED STATES DEPARTMENT OF  
COMMERCE; GINA RAIMONDO, in her  
official capacity as Secretary of Commerce;  
UNITED STATES BUREAU OF THE  
CENSUS, an agency within the United States  
Department of Commerce; and RON  
JARMIN, in his official capacity as Acting  
Director of the U.S. Census Bureau,

Defendants.

CASE NO. 3:21-cv-211-RAH

**DECLARATION OF DR. MICHAEL BARBER**

MICHAEL BARBER pursuant to 28 U.S.C. § 1746, Federal Rule of Civil Procedure

26(a)(2)(B), and Rules 702 and 703 of the Federal Rules of Evidence, declares as follows:

1. I am 37 years old and competent to make this declaration.
2. I am an associate professor of political science at Brigham Young University and faculty fellow at the Center for the Study of Elections and Democracy in Provo, Utah.
3. I received a PhD in politics from Princeton University in 2014 with emphases in American politics and quantitative methods/statistical analyses.
4. I teach a number of undergraduate courses in American politics and quantitative research methods.
5. I have over 10 years of experience conducting complex demographic and analytical analyses.
6. My research focus is on election and voting related topics in American politics and public opinion. Much of my research uses advanced statistical methods for the analysis of quantitative data including census data.
7. I have previously qualified as an expert witness in federal court.
8. Plaintiffs requested that I assess the Census Bureau's disclosure avoidance techniques including differential privacy, and to describe the mechanisms by which differential privacy works and the resulting impacts on the end users of census data.
9. I am being compensated \$400 an hour for my time in connection with this matter. I am not being compensated for any specific opinion.
10. Attached and incorporated by reference to this declaration is my expert report in this matter and my curriculum vitae. The report is attached hereto as Appendix 1. My curriculum vitae is attached to the expert report as Appendix A.

11. My curriculum vitae lists, among other things, my qualifications, a list of all publications published over at least the last ten years, and a list of all cases over at least the past four years in which I testified as an expert at trial or by deposition.

12. I declare under penalty of perjury that the foregoing, including any appendices, are true and correct according to the best of my knowledge, information, and belief.

Dated: March 9, 2021



Michael Barber

# Appendix 1



## Expert Report of Michael Barber

Dr. Michael Barber  
Brigham Young University  
724 Spencer W. Kimball Tower  
Provo, UT 84604  
barber@byu.edu

# 1 Introduction and Qualifications

I am an associate professor of political science at Brigham Young University and faculty fellow at the Center for the Study of Elections and Democracy in Provo, Utah. I received my PhD in political science from Princeton University in 2014 with emphases in American politics and quantitative methods/statistical analyses. My dissertation was awarded the 2014 Carl Albert Award for best dissertation in the area of American Politics by the American Political Science Association.

I teach a number of undergraduate courses in American politics and quantitative research methods.<sup>1</sup> These include classes about political representation, Congressional elections, statistical methods, and research design.

I have worked as an expert witness in a number of cases in which I have been asked to perform and evaluate various statistical methods. Cases in which I have testified at trial or by deposition are listed in my CV, which is attached to the end of this report.

In my position as a professor of political science, I have conducted research on a variety of election- and voting-related topics in American politics and public opinion. Much of my research uses advanced statistical methods for the analysis of quantitative data. I have worked on a number of research projects that use “big data” that include millions of observations, including a number of state voter files, campaign contribution lists, and data from the US Census.

Much of this research has been published in peer-reviewed journals. I have published nearly 20 peer-reviewed articles, including in our discipline’s flagship journal, *The American Political Science Review* as well as the inter-disciplinary journal, *Science Advances*. My CV, which details my complete publication record, is attached to this report as Appendix A.

The analysis and explanation I provide in this report are consistent with my training in statistical analysis and are well-suited for this type of analysis in political science and quantitative analysis more generally.

---

<sup>1</sup>The political science department at Brigham Young University does not offer any graduate degrees.

I teach a number of undergraduate courses in American politics and quantitative research methods.<sup>2</sup> These include classes about political representation, Congressional elections, statistical methods, and research design.

I have worked as an expert witness in a number of cases where I have been asked to evaluate and perform various statistical analyses. Cases in which I have testified at trial or by deposition are listed in my CV, which is attached to the end of this report.

I have been asked to evaluate and explain at an approachable level the process of differential privacy (DP), its application to the 2020 Census, and how it fits within the field of probability theory and statistical methods.

## 2 Introduction to Statistical Disclosure Limitations

The Census collects the confidential information of Americans under Title 13 of the U.S. Code. 13 U.S. Code 9 requires that the confidentiality of these individuals' records be protected and prohibits the Census from making "any publication whereby the data furnished by any particular establishment or individual under this title can be identified." In other words, it should not be possible for a person using the aggregate tables published by the Census Bureau to use those data to identify specific individuals.

To protect individuals' identity, in some cases it becomes necessary in some cases to alter the original data because a person living in a particular area may be unique enough to be identified. For example, an exceptionally wealthy person or an individual who is the only member of a particular race in their census block might be identifiable even when using aggregate statistics.

There are a variety of approaches to accomplish statistical disclosure limitation (SDL), many of which have been used by Census in the past. Aboud, et al (2020) discusses these different approaches: "Historically, the Census Bureau has primarily used information reduction and data perturbation methods to support SDL (Lauger et al., 2014). Information

---

<sup>2</sup>The political science department at Brigham Young University does not offer any graduate degrees.

reduction methods include top- and bottom-coding, suppression, rounding or binning, and sampling collected units for release in public use microdata files. Data perturbation methods include swapping, legacy noise injection systems, and partially and fully synthetic database construction. These legacy approaches start with the premise that there are specific data elements that must be protected (e.g., a person's income). A technical analyst chooses an approach from the assortment of available SDL methods that is likely to protect the data without resulting in too much damage to the published data accuracy. Usually, the selection of SDL method takes into consideration the intended uses of the published data along with assumptions about the kind of external data an intruder might have, and the types of privacy attacks an intruder might attempt.”<sup>3</sup>

There is a vast literature of scholarly research on these and many other methods of SDL. Wasserman and Zhou (2010), Reiter (2018), and Karr (2016) all provide an excellent summary of many of these methods as well as associated scholarly research in computer science and statistics regarding these approaches.<sup>4</sup>

For example, Karr (2016) provides a classification of various SDL methods, many of which have been used by the Census Bureau in the past: “There are three principal classes of SDL methods. The first class is reduction techniques that do not alter data values. These include cell suppression, subsampling, variable deletion, top-coding, bottom-coding, category aggregation, and conversion of numerical variables to categorical ranges (Kinney et al. 2009). The second class is perturbative methods such as addition of noise, micro-aggregation, and data swapping, as well as combinations of methods (Oganian & Karr 2006, Singh 2010). The

---

<sup>3</sup><https://www2.census.gov/adrm/CED/Papers/CY20/2020\protect\discretionary{\char\hyphenchar\font}{-}{08\protect\discretionary{\char\hyphenchar\font}{-}{-}AbowdBenedettoGarfinkelDahleta\protect\discretionary{\char\hyphenchar\font}{-}{-}The%20modernization%20of.pdf>  
Lauger, Amy, Billy Wisniewski, and Laura McKenna (2014). Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research. Research Report Series (Disclosure Avoidance #2014-02). Washington: Center for Disclosure Avoidance Research, U.S. Census Bureau.

<sup>4</sup>Larry Wasserman & Shuheng Zhou (2010) A Statistical Framework for Differential Privacy, Journal of the American Statistical Association, 105:489, 375-389, DOI: 10.1198/jasa.2009.tm08651  
Reiter, Jerome P. “Differential privacy and federal data releases.” Annual review of statistics and its application 6 (2019): 85-101  
Karr, Alan F. “Data sharing and access.” Annual Review of Statistics and Its Application 3 (2016): 113-132.

third class is synthetic data methods originating from techniques for imputation of missing data, in which some, or in extreme cases all, variables are replaced by values generated by a Bayesian posterior predictive distribution (Reiter 2005a,b,c; Reiter et al. 2014).”<sup>5</sup>

The National Conference of State Legislatures, an organization that works with, and advocates on behalf of, state legislatures around the country, discusses this issue and provides an excellent example. “Consider a census block with just 20 people in it, including one Filipino American. Without any disclosure avoidance effort, it might be possible to figure out the identity of that individual. With data swapping, the Filipino American’s data might be swapped with that of an Anglo American from a nearby census block—a census block where other Filipino Americans reside. The details for the person would be aggregated with others, and therefore not identifiable, and yet the total population in both census blocks would remain accurate.”<sup>6</sup>

### 3 Differential Privacy in the 2020 Census

In the 2020 Census, in addition to many of the SDL methods used in previous decades, the Census Bureau plans to also introduce the concept of differential privacy.<sup>7</sup> Differential

---

<sup>5</sup>Karr, Alan F. “Data sharing and access.” *Annual Review of Statistics and Its Application* 3 (2016): 113-132

Kinney SK, Gonzalez JF Jr, Karr AF. 2009. Data confidentiality—the next five years: summary and guide to papers. *J. Priv. Confid.* 1(2):125–34

Oganian A, Karr AF. 2006. Combinations of SDC methods for microdata protection. In *Privacy in Statistical Databases*, ed. J Domingo-Ferrer, L Franconi, pp. 102–13. *Lect. Notes Comput. Sci. Ser.* 4302. New York: Springer-Verlag, Singh AC. 2010. Maintaining analytic utility while protecting confidentiality of survey and nonsurvey data. *J. Priv. Confid.* 1(2):155–82

Reiter JP. 2005a. Estimating risks of identification disclosure for microdata. *J. Am. Stat. Assoc.* 100:1103–13

Reiter JP. 2005b. Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. *J. R. Stat. Soc. Ser. A* 168:185–205

Reiter JP. 2005c. Using CART to generate partially synthetic, public use microdata. *J. Off. Stat.* 21:441–62

<sup>6</sup><https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx>

<sup>7</sup><https://www2.census.gov/ces/wp/2018/CES-WP-18-47.pdf>

<https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>

<https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products/faqs.html>

privacy is a particular type of SDL and is a relatively new application of statistical methods, having been developed within the last 20 years.<sup>8</sup> Differential privacy uses statistical distributions to alter the data by allocating a pre-determined “privacy budget” across different levels of data. These alterations make it increasingly difficult to identify an individual’s record in the data. The Harvard University Privacy Tools Project provides a way to think conceptually about differential privacy as a matter of the probability of an individual’s information being revealed not substantially increasing if their information is contained in a database. They state: “Consider an algorithm that analyzes a dataset and computes statistics about it (such as the data’s mean, variance, median, mode, etc.). Such an algorithm is said to be differentially private if by looking at the output, one cannot tell whether any individual’s data was included in the original dataset or not. In other words, the guarantee of a differentially private algorithm is that its behavior hardly changes when a single individual joins or leaves the dataset.”<sup>9</sup>

To accomplish this, differential privacy uses various statistical methods to alter (perturb or distort) the original dataset in such a way so as to make it less possible to infer the identity of any individual by looking at any part of the distorted dataset – whether individual records or summary statistics. The JASON advisory group, an independent group of scientists which advise the United States government on matters of science and technology, was asked to review the Census Bureau’s plan to use differential privacy in 2020. Their report summarizes the process well. They state that differential privacy “makes possible statistical queries regarding a dataset to be performed while offering a rigorous bound on the amount one learns about a dataset if one record is deleted, added or replaced. Note that this is not, strictly speaking a guarantee of disclosure avoidance, but it does provide in a rigorous way the likelihood of a record linkage attack. It does this by adding specially calibrated noise

---

Ashmead, Robert, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. *Effective Privacy After Adjusting for Invariants with Applications to the 2020 Census*. Technical Report. US Census Bureau, 2019.

<sup>8</sup>Hilton, Michael. “Differential privacy: a historical survey.” Cal Poly State University (2002).

<sup>9</sup><https://privacytools.seas.harvard.edu/differential-privacy>

to the result of a specific query made on the dataset...The value set for the privacy loss parameter is meant to be a policy decision.”<sup>10</sup>

Using the pre-determined privacy budget in conjunction with the chosen statistical distribution, the researcher can in essence “dial up and down” the degree of privacy (typically noted as the Greek letter epsilon) by increasing or decreasing the level to which the original data are altered via parameters set in the statistical distributions used in the method.<sup>11</sup> “Accepted guidelines for choosing epsilon have not yet been developed....The exact choice of epsilon is a policy decision that should depend on the sensitivity of the data, with whom the output will be shared, the intended data analysts’ accuracy requirements, and other technical and normative factors.”<sup>12</sup>

As a part of its implementation, differential privacy requires a number of decisions and inputs from the researcher. First, there are a variety of different methods by which a researcher can implement differential privacy. One such example is the choice of statistical distribution and the parameters set in that distribution to introduce “noise” for each record in the database. For example, two common distributions that have been used are the Geometric distribution and the Laplace distribution.<sup>13</sup> These distributions are commonly used in various applications of statistics and probability theory. In the context of differential privacy, the process occurs in two steps. First parameters are chosen to calibrate the variance of the chosen distribution, and then random draws from these distributions are taken and applied to the observations to “perturb” or “alter” values in the database up or down by adding the value of the random draw, which can be either positive or negative.

The decision of parameter values in these distributions is made by the researcher as

---

<sup>10</sup><https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/privacy-methods-2020-census.html>, pg. 14

<sup>11</sup>Reiter, Jerome P. “Differential privacy and federal data releases.” Annual review of statistics and its application 6 (2019): 85-101

<sup>12</sup>Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, and Salil Vadhan. “Differential privacy: A primer for a non-technical audience.” Vanderbilt Journal of Entertainment & Technology Law 21, no. 1 (2018): 209-275.

<sup>13</sup>Abowd, John, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. Technical Report. US Census Bureau, 2019.

he or she decides how much accuracy to retain from the original, unperturbed database, and how much privacy to introduce into the altered database by obscuring the original values. This tradeoff between accuracy and privacy is known as the “risk-utility paradigm” since greater accuracy increases the risk of identification of individual records while greater privacy decreases the utility of the distorted database since the values in each record (or summary statistics based on those perturbed individual records) become less accurate.

For example, consider the extremes of the “risk-utility” continuum. Making no changes (i.e. implementing no privacy measures) puts individuals at the greatest risk of identification but also provides researchers the greatest utility since they know that the statistics they calculate from the data are based on an entirely accurate database (or at least accurate insofar as the data have been collected accurately and have not been altered by whomever collected the data). However, no additional privacy is afforded those individuals whose information is contained in the database. At the other extreme, individuals’ privacy can be absolutely guaranteed if no information at all is made available, or if the information is altered so greatly as to be entirely worthless. This affords perfect privacy; however, the database has no utility to researchers or policymakers.

Karr (2016) summarizes this tradeoff by stating: “Modern approaches to SDL are based explicitly or implicitly on a trade-off between disclosure risk and data utility (Cox et al. 2011). Crucially, higher risk and higher utility go together. No release, the only action that means no risk, also means no utility to analysts. The risk-utility approach requires quantified measures of both disclosure risk and data utility for each candidate SDL method and setting of parameters within it.”<sup>14</sup>

In the context of the US Census, the discussion of risk in the risk-utility paradigm would include all individuals whose information is contained in the Census records while the discussion “utility” in the risk-utility paradigm would include government agencies, state

---

<sup>14</sup>Data Sharing and Access, Alan F. Karr, *Annu. Rev. Stat. Appl.* 2016. 3:113–32  
Cox LH, Karr AF, Kinney SK. 2011. Risk-utility paradigms for statistical disclosure limitation: how to think, but not how to act (with discussion). *Int. Stat. Rev.* 79(2):160–99



legislatures, interest groups, scholars, and other policymaking organizations who regularly use and rely upon summary statistics of the population derived from the decennial Census data to guide their decision-making, research, and advocacy. The National Conference of State Legislatures discusses differential privacy in the 2020 Census and how it may affect policies and procedures taken up by various state legislatures. “Differential privacy will mean that, except at the state level, population and voting age population will not be reported as enumerated. And, race and ethnicity data are likely to be farther from the “as enumerated” data than in past decades, when data swapping was used to protect small populations. (In 2010, at the block level, total population, voting age population, total housing units, occupancy status, group quarters count and group quarters type were all held invariant.) This may raise issues for racial block voting analyses.”<sup>15</sup>

Individual states have also expressed their concern that if the degree of privacy in the 2020 Census is set too far towards the privacy side of the risk-utility scale, it may have negative effects with regards to policymaking, legislative redistricting, the allocation of government funding, or simply having an accurate measure of the state of affairs in their states and municipalities.<sup>16</sup>

While the mathematical and statistical details of the algorithms used to implement differential privacy in the 2020 Census are computationally intense and highly technical, the overall process can be described in a general sense quite simply. I omit the technical details here, but they are contained in various documents published by Census researchers.<sup>17</sup> The

<sup>15</sup><https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx>

<sup>16</sup>[https://www.ncsl.org/Portals/1/Documents/Redistricting/VA\\_CensusDistortionProgram\\_VAGovernor\\_2020-01-23.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/VA_CensusDistortionProgram_VAGovernor_2020-01-23.pdf)  
[https://www.ncsl.org/Portals/1/Documents/Redistricting/WA\\_OFM\\_DAS\\_Response\\_Letter.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/WA_OFM_DAS_Response_Letter.pdf)  
[https://www.ncsl.org/Portals/1/Documents/Redistricting/UT\\_Differential\\_Privacy\\_%28Signed%29.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/UT_Differential_Privacy_%28Signed%29.pdf)

<sup>17</sup>Abowd, John, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. Technical Report. US Census Bureau, 2019.  
[https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711\\_0938\\_2018\\_E2E\\_Test\\_Algorithm\\_Description.pdf](https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0938_2018_E2E_Test_Algorithm_Description.pdf)  
[https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711\\_0941\\_Effective\\_Privacy\\_after\\_Adjusting\\_for\\_Constraints\\_With\\_applications\\_to\\_the\\_2020\\_Census.pdf](https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0941_Effective_Privacy_after_Adjusting_for_Constraints_With_applications_to_the_2020_Census.pdf)

process is referred to as “TopDown”, as it begins with the entire country and subsequently applies privacy measures to lower and lower geographies (i.e. state, county, tract, block group, block).

The first step is to create a multi-dimensional histogram based on the intersection of the variables collected in the Census in the PL94-171 microdata – a table with attributes Race (63 possible values), Ethnicity (Hispanic or not), Voting Age (whether age is 18+ or not), and Housing Type (nine possible values), and location (state, county, tract, block group, and block). The intersection of these variables would create an enormously large set of cells, particularly given the number of unique blocks in the country (i.e. the unique intersection of each variable and geographic unit, such as the number of White, non-Hispanic, and 18+ persons living in single family dwelling units in a particular census block would be computationally intractable). As such, the PL94-171 dataset is too large to process at once, and so the TopDown algorithm begins by creating a national histogram for the entire country, leaving out the various smaller geographic units.

The algorithm then samples from a statistical distribution (the Geometric or Laplace distribution) with parameters set to the desired level of variance (higher variance yields greater noise injection and thus greater privacy and less accuracy) and applies this perturbation to the values in each cell. It then solves a minimization procedure (for example, least squares) to select the optimal “noisy” histogram.

The degree of noise injected via the statistical sampling from the Geometric or Laplace distributions is a direct consequence of a choice made by the Census Bureau regarding the degree of privacy that should be present versus the degree of accuracy that should remain in the altered database (the choice of epsilon). Riper, Kugler, and Ruggles (2020) describe this process in the following way: “The global privacy-loss budget (PLB), usually denoted by the Greek letter epsilon, establishes the trade-off between the privacy afforded to Census respondents and the accuracy of the published data. Values for epsilon range from essentially

---

pdf

0 to infinity, with 0 representing perfect privacy/no accuracy and infinity representing no privacy/perfect accuracy. Once the global PLB is established, it can then be spent by allocating fractions to particular geographic levels and queries. Geographic levels or queries that receive larger fractions will be more accurate, and levels or queries that receive smaller fractions or no specific allocation will be less accurate (pg. 357).<sup>18</sup>

The application of this random noise will in some cases cause cells to extend beyond logical values (i.e. cells with negative numbers), and the total sum of the cells must still sum to the actual total in the population. Thus, the “noisy” histogram is further adjusted to constrain cells to meet these criteria. Finally, the histogram values are constrained to be integer values (i.e. whole numbers - no fractions of people) while the sum of the cells must still sum to the total population. This means that there will not be any blocks with -3 or 5.35 people in them.

This process is then repeated down the geographic “spine” of the census. “This process happens recursively—first, we fix (i.e., hold constant) the root node and generate its children (e.g., histograms for each state) with the constraint that the child histograms add up to the parent histogram while satisfying their own implied constraints. Then, for each state histogram, we fix the histogram and generate its county-level children such that they add up to the state, and so forth down to the block (pg. 7).<sup>19</sup>

The final constraint is the introduction of “invariants”, or statistics for which the Census Bureau has committed to providing the exact values rather than the statistically altered, noisy values. For the 2020 Census, the Census Bureau currently plans to provide the following invariants: total number of people per state, total number of housing units by block, and number of group quarters facilities by block.<sup>20</sup> These same four counts were

<sup>18</sup>Van Riper, David, Tracy Kugler, and Steven Ruggles. “Disclosure Avoidance in the Census Bureau’s 2010 Demonstration Data Product.” In *International Conference on Privacy in Statistical Databases*, pp. 353-368. Springer, Cham, 2020.

<sup>19</sup>Abowd, John, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. *Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge*. Technical Report. US Census Bureau, 2019.

<sup>20</sup><https://content.govdelivery.com/accounts/USCENSUS/bulletins/2ae5eda>

invariant at the census block-level (note the difference in 2010 where total population is invariant at the block rather than state level) in the 2010 Decennial Census. Additionally, voting age population and occupied housing units (i.e., households) were invariant at the census block-level in 2010.<sup>21</sup> The final product is a new database that has these statistically altered values that protect the privacy of those included in the original database but also contains the accurate values for the variables determined to be held invariant.

The lack of invariant populations at the block level (as was included in the 2010 Census data) poses significant issues for state legislatures and other bodies tasked with the creation of legislative districts that are required by law to contain equal populations. As redistricting bodies assemble districts by drawing lines across their respective states, they depend on accurate population data to ensure that those districts contain equal populations. Moreover, in some cases districts are designed to contain certain percentages of minority populations, which becomes increasingly difficult without accurate counts. Legislative leaders in the state of Utah expressed concerns similar to this in a letter to Census Director Dillingham in early 2020: “[W]ith respect to redistricting, we notice larger population shifts than expected, particularly within legislative house districts. Consequently, we fear that differential privacy will require the states to legally defend whether differential privacy protected census data will satisfy the states’ constitutional obligation to meet population and equality requirements. Based upon our analysis of differential privacy as applied to the 2010 census redistricting data, we believe, if differential privacy is applied to the 2020 redistricting data, that the integrity of the data used to redistrict the state into congressional and legislative districts, and also within our local jurisdictions, will be threatened.”<sup>22</sup>

---

<sup>21</sup>Van Riper, David, Tracy Kugler, and Steven Ruggles. “Disclosure Avoidance in the Census Bureau’s 2010 Demonstration Data Product.” In International Conference on Privacy in Statistical Databases, pp. 353-368. Springer, Cham, 2020.

<sup>22</sup>[https://www.ncsl.org/Portals/1/Documents/Redistricting/UT\\_Differential\\_Privacy\\_%28Signed%29.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/UT_Differential_Privacy_%28Signed%29.pdf)

## 4 Inequitable Distribution of Intentionally Introduced Error

Riper, Kugler, and Ruggles (2020) also note that the implementation of the non-negativity and block-level total housing unit invariant constraints can lead to greater error as well as error that is biased in a particular direction in the new “altered” database. “The non-negativity constraint requires that every cell in the final detailed histogram be non-negative. As described above, many of the cells in the noisy household histograms will be negative, especially for geographic units with smaller numbers of households. Returning these cells to zero effectively adds households to these small places, resulting in positive bias. . . The invariant number of housing units down to the block level implies an upper-bound constraint on the number of households. Each geographic unit must have no more households than it has housing units. With the low signal-to-noise ratio in the noisy histograms, especially at the block level, this constraint is the strongest signal present in the optimization problem. Many geographic units therefore receive a number of households equal to the number of housing units, resulting in 100% occupancy rates. This is especially true for geographic units with smaller numbers of households that are affected by positive bias due to the non-negativity constraint...The issue of scale-independent noise affects all of the millions of cells with small counts in both the person and household histograms, making counts of many population subsets unreliable. The combination of the non-negativity constraint and population invariants consistently leads to bias increasing counts of small subgroups and small geographic units and decreasing counts of larger subgroups and geographic units. (pg 363-364).”

As noted earlier, the process of differential privacy is not applied equally across the entire population. Places with fewer people (rural locations) and areas with smaller, distinctive populations (minority communities) are more likely to be impacted since these are the places where identification is more concerning, and the application of statistical noise

is more likely to have a larger impact on the summary statistics derived from the altered data. This is especially the case when reported statistics must be in whole numbers (i.e. no fractional people or housing units). A simplified example helps illustrate the point. Suppose there are two census blocks, one with 10 people and another with 100 people. The block with 10 people is more susceptible to an identification “attack” given its smaller population. Furthermore, any perturbations will have a larger impact on the summary statistics of the smaller block — a change in the ethnicity of one individual in the smaller block represents a 10% change overall while one individual change in the large block represents a 1% alteration to the summary statistics in the block. Furthermore, to add noise to small blocks without having negative population numbers requires that small blocks, on average, get bigger. In turn, because of the decision to keep state population invariant (i.e. accurate), this means that the largest blocks, on average, get smaller.<sup>23</sup>

Garfinkle, et al (2018) succinctly summarize the situation: “By design, the noise-injection mechanisms used by the Census Bureau will result in increased accuracy as population sizes increase.”<sup>24</sup> Santos-Lozada, et al (2020) elaborate on this issue and discuss some of the potential problems it presents: “Infusing noise in the data, in comparison to the current disclosure avoidance system, will produce inaccurate patterns of demographic change with higher levels of error found in the calculations for non-Hispanic blacks and Hispanics. At the same time, these counts are bound to impact post-2020 districting for both federal and state elections, as well as evaluations of that redistricting. . . . [T]hese changes in population counts will affect understandings of health disparities in the nation, leading to overestimates of population-level health metrics of minority populations in smaller areas and underestimates of mortality levels in more populated ones.”<sup>25</sup> Pujol, et al (2020) provide a generalized study of how the application of differential privacy may “disproportionately

<sup>23</sup>[https://datasociety.net/wp-content/uploads/2019/12/DS\\_Differential\\_Privacy\\_L.pdf](https://datasociety.net/wp-content/uploads/2019/12/DS_Differential_Privacy_L.pdf)

<sup>24</sup>Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. “Issues encountered deploying differential privacy.” In Proceedings of the 2018 Workshop on Privacy in the Electronic Society, pp. 133-137. 2018.

<sup>25</sup>Santos-Lozada, Alexis R., Jeffrey T. Howard, and Ashton M. Verdery. “How differential privacy will affect our understanding of health disparities in the United States.” Proceedings of the National Academy of Sciences 117, no. 24 (2020): 13405-13412.

impact some groups over others (pg. 189)” and find disparities among smaller populations with important applications to the allocation of government funds, voting rights benefits to minority communities, and calculations for the apportionment of legislative seats.<sup>26</sup>

Individual states have also taken note of the potential problems that this issue can present. For example, the state of Washington, when testing the proposed 2020 process on 2010 census data found, “There is a bias in the demonstration data that causes areas with small populations to get larger while areas with larger populations get smaller.” They also found, “There is another bias in the data that makes communities with similar racial characteristics more dispersed geographically.”<sup>27</sup> The state of Utah came to similar conclusions when looking at the test data in their state. They find, “We observe that the population loss in our cities and towns are re-allocated to unincorporated, rural areas of the state,” and that “we are currently assessing how this net loss will impact state and federal funding that is disbursed in compliance with state revenue sharing statutes and federally mandated population formulas.”<sup>28</sup> California has expressed concern regarding the application of differential privacy with regard to the number of persons residing in census blocks containing incarcerated individuals.<sup>29</sup>

In July 2020, Census Director Dillingham wrote to the National Conference of State Legislatures to address their concerns of how differential privacy might impact the PL94-171 redistricting data. In his letter, he noted that after an adjustment to the “operations in the post processing algorithms” in the 2010 demonstration data products provided by the Census Bureau that there were notable “improvements in accuracy for total population counts.” However, there are still large differences, particularly at smaller levels of geography. For example, “At the block level, error in the population for the average urban census

<sup>26</sup>Pujol, David, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. “Fair decision making using privacy-protected data.” In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 189-199. 2020.

<sup>27</sup>[https://www.ncsl.org/Portals/1/Documents/Redistricting/WA\\_OFM\\_DAS\\_Response\\_Letter.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/WA_OFM_DAS_Response_Letter.pdf)

<sup>28</sup>[https://www.ncsl.org/Portals/1/Documents/Redistricting/UT\\_Differential\\_Privacy\\_%28Signed%29.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/UT_Differential_Privacy_%28Signed%29.pdf)

<sup>29</sup>[https://www.ncsl.org/Portals/1/Documents/Redistricting/California\\_Leaders\\_Letter\\_to\\_RonaldKlain\\_Feb2021.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/California_Leaders_Letter_to_RonaldKlain_Feb2021.pdf)

block has likewise been reduced from 9.2 people to 7.7 people.”<sup>30</sup> The difficulty faced by policymakers is that 7.7 people can represent a substantial proportion of a census block, given their extremely small size. While there is certainly variation, a simple calculation of the total US population divided by the number of census blocks yields an average population per block of 28 people. An average variation of 7.7 people would represent an average error of more than 25%.<sup>31</sup> The metrics tables released with the 2010 demonstration data indicate that even with this reduction in error, nearly 50 percent of blocks classified as “urban” contained an error larger than 5% while 36% of blocks classified as “rural” contained an error larger than 5%.<sup>32</sup> These differences could pose significant problems for states such as Alabama that are trying to satisfy legal requirements of one person, one vote or the creation of majority-minority districts in their redistricting process.

## 5 Differential Privacy in the Context of Probability and Statistics

While the idea of differential privacy has its roots in computer science, the procedure can be thought of as a question of probability theory and statistical methods. At its core, the process of ensuring privacy is a combination of sampling and constrained optimization. Privacy is introduced into the data by introducing random error through sampling from statistical distributions with parameters set to a desired level of variance (privacy). These random draws are then added or subtracted to the actual observations (or summary statistics

---

<sup>30</sup>[https://www.ncsl.org/Portals/1/Documents/Elections/CensusBureau\\_letter\\_to\\_NCSL%20Storey\\_071620.pdf](https://www.ncsl.org/Portals/1/Documents/Elections/CensusBureau_letter_to_NCSL%20Storey_071620.pdf)

<sup>31</sup>To produce the average population per block I simply took the total national population as of the 2010 census (308,746,065) and divided it by the total number of blocks (11,078,300) in the 2010 census. However, many blocks are uninhabited (4,871,270). Removing them from the calculation would yield an average population per block of approximately 50 people and an average error of  $7.7/50 = .15$ . See <https://tumblr.mapsbynik.com/post/82791188950/nobody-lives-here-the-4-million-census-blocks> for total number of blocks and unpopulated blocks.

<sup>32</sup><https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>, see “Detailed Summary Metrics” file.



of the actual observations in the form of histograms) and then further adjusted to conform to certain constraints using an algorithm that is designed to minimize the alterations necessary to achieve this objective. The final process is akin to a constrained optimization problem, which is well within the wheelhouse of statistics and econometrics. For a more detailed and mathematical consideration of the relationship between differential privacy and probability theory and statistical inference, see Wasserman and Zhou (2010) and Dwork and Smith (2010).<sup>33</sup> Differential privacy is thus an application of statistical processes and methods to adjust the original counts of the Census to protect the privacy of individual's records. It is dramatically different in its methods and application from the methods used previously to protect the identity of individuals in the Census.

---

<sup>33</sup>Larry Wasserman & Shuheng Zhou (2010) A Statistical Framework for Differential Privacy, *Journal of the American Statistical Association*, 105:489, 375-389, DOI: 10.1198/jasa.2009.tm08651  
Dwork, Cynthia, and Adam Smith. "Differential privacy for statistics: What we know and what we want to learn." *Journal of Privacy and Confidentiality* 1, no. 2 (2010).

## Appendix A - Curriculum Vitae

# Michael Jay Barber

---

## CONTACT INFORMATION

Brigham Young University  
Department of Political Science  
724 KMBL  
Provo, UT 84602

barber@byu.edu  
<http://michaeljaybarber.com>  
Ph: (801) 422-7492

## ACADEMIC APPOINTMENTS

**Brigham Young University, Provo, UT**

- 2020 - present Associate Professor, Department of Political Science
- 2014 - 2020 Assistant Professor, Department of Political Science
  
- 2014 - present Faculty Scholar, Center for the Study of Elections and Democracy

## EDUCATION

**Princeton University Department of Politics, Princeton, NJ**

Ph.D., Politics, July 2014

- Advisors: Brandice Canes-Wrone, Nolan McCarty, and Kosuke Imai
- Dissertation: "Buying Representation: the Incentives, Ideology, and Influence of Campaign Contributions on American Politics"
- 2015 Carl Albert Award for Best Dissertation, Legislative Studies Section, American Political Science Association (APSA)

M.A., Politics, December 2011

**Brigham Young University, Provo, UT**

B.A., International Relations - Political Economy Focus, April, 2008

- *Cum Laude*

## RESEARCH INTERESTS

American politics, congressional polarization, political ideology, campaign finance, survey research

## PUBLICATIONS

18. "Comparing Campaign Finance and Vote Based Measures of Ideology"  
Forthcoming at *Journal of Politics*
17. "The Participatory and Partisan Impacts of Mandatory Vote-by-Mail", with John Holbein  
*Science Advances*, 2020. Vol. 6, no. 35, DOI: 10.1126/sciadv.abc7685
16. "Issue Politicization and Interest Group Campaign Contribution Strategies", with Mandi Eatough  
*Journal of Politics*, 2020. Vol. 82: No. 3, pp. 1008-1025
15. "Campaign Contributions and Donors' Policy Agreement with Presidential Candidates", with Brandice Canes-Wrone and Sharece Thrower  
*Presidential Studies Quarterly*, 2019, 49 (4) 770-797

14. **“Conservatism in the Era of Trump”**, with Jeremy Pope  
*Perspectives on Politics*, 2019, 17 (3) 719–736
13. **“Legislative Constraints on Executive Unilateralism in Separation of Powers Systems”**, with Alex Bolton and Sharece Thrower  
*Legislative Studies Quarterly*, 2019, 44 (3) 515–548  
Awarded the Jewell-Loewenberg Award for best article in the area of subnational politics published in *Legislative Studies Quarterly* in 2019
12. **“Electoral Competitiveness and Legislative Productivity”**, with Soren Schmidt  
*American Politics Research*, 2019, 47 (4) 683–708
11. **“Does Party Trump Ideology? Disentangling Party and Ideology in America”**, with Jeremy Pope  
*American Political Science Review*, 2019, 113 (1) 38–54
10. **“The Evolution of National Constitutions”**, with Scott Abramson  
*Quarterly Journal of Political Science*, 2019, 14 (1) 89–114
9. **“Who is Ideological? Measuring Ideological Responses to Policy Questions in the American Public”**, with Jeremy Pope  
*The Forum: A Journal of Applied Research in Contemporary Politics*, 2018, 16 (1) 97–122
8. **“Status Quo Bias in Ballot Wording”**, with David Gordon, Ryan Hill, and Joe Price  
*The Journal of Experimental Political Science*, 2017, 4 (2) 151–160.
7. **“Ideologically Sophisticated Donors: Which Candidates Do Individual Contributors Finance?”**, with Brandice Canes-Wrone and Sharece Thrower  
*American Journal of Political Science*, 2017, 61 (2) 271–288.
6. **“Gender Inequalities in Campaign Finance: A Regression Discontinuity Design”**, with Daniel Butler and Jessica Preece  
*Quarterly Journal of Political Science*, 2016, Vol. 11, No. 2: 219–248.
5. **“Representing the Preferences of Donors, Partisans, and Voters in the U.S. Senate”**  
*Public Opinion Quarterly*, 2016, 80: 225–249.
4. **“Donation Motivations: Testing Theories of Access and Ideology”**  
*Political Research Quarterly*, 2016, 69 (1) 148–160.
3. **“Ideological Donors, Contribution Limits, and the Polarization of State Legislatures”**  
*Journal of Politics*, 2016, 78 (1) 296–310.
2. **“Online Polls and Registration Based Sampling: A New Method for Pre-Election Polling”** with Quin Monson, Kelly Patterson and Chris Mann.  
*Political Analysis* 2014, 22 (3) 321–335.
1. **“Causes and Consequences of Political Polarization”** In *Negotiating Agreement in Politics*. Jane Mansbridge and Cathie Jo Martin, eds., Washington, DC: American Political Science Association: 19–53. with Nolan McCarty. 2013.
  - Reprinted in *Solutions to Political Polarization in America*, Cambridge University Press. Nate Persily, eds. 2015
  - Reprinted in *Political Negotiation: A Handbook*, Brookings Institution Press. Jane Mansbridge and Cathie Jo Martin, eds. 2015

AVAILABLE  
WORKING PAPERS

**“Ideological Disagreement and Pre-emption in Municipal Policymaking”**  
with Adam Dynes (Revise and Resubmit at *American Journal of Political Science*)

**“Taking Cues When You Don’t Care: Issue Importance and Partisan Cue Taking”**  
with Jeremy Pope (Revise and Resubmit at *Public Opinion Quarterly*)

**“A Revolution of Rights in American Founding Documents”**  
with Scott Abramson and Jeremy Pope (Under Review)

**“410 Million Voting Records Show That Minority Citizens, Young People, and Democrats Are at a Profound Disadvantage at the Ballot Box”**  
with John Holbein (Under Review)

**“Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records”** (Under Review)

**“Partisanship and Trolleyology”**  
with Ryan Davis (Under Review)

**“Who’s the Partisan: Are Issues or Groups More Important to Partisanship?”**  
with Jeremy Pope (Under Review)

**“The Policy Preferences of Donors and Voters”**

**“Estimating Neighborhood Effects on Turnout from Geocoded Voter Registration Records.”**  
with Kosuke Imai

**“Super PAC Contributions in Congressional Elections”**

WORKS IN  
PROGRESS

**“Collaborative Study of Democracy and Politics”**  
with Brandice Canes-Wrone, Gregory Huber, and Joshua Clinton

**“Preferences for Representational Styles in the American Public”**  
with Ryan Davis and Adam Dynes

**“Representation and Issue Congruence in Congress”**  
with Taylor Petersen

**“Education, Income, and the Vote for Trump”**  
with Edie Ellison

INVITED  
PRESENTATIONS

**“Are Mormons Breaking Up with Republicanism? The Unique Political Behavior of Mormons in the 2016 Presidential Election”**

- Ivy League LDS Student Association Conference - Princeton University, November 2018, Princeton, NJ

**“Issue Politicization and Access-Oriented Giving: A Theory of PAC Contribution Behavior”**

- Vanderbilt University, May 2017, Nashville, TN

“Lost in Issue Space? Measuring Levels of Ideology in the American Public”

- Yale University, April 2016, New Haven, CT

“The Incentives, Ideology, and Influence of Campaign Donors in American Politics”

- University of Oklahoma, April 2016, Norman, OK

“Lost in Issue Space? Measuring Levels of Ideology in the American Public”

- University of Wisconsin - Madison, February 2016, Madison, WI

“Polarization and Campaign Contributors: Motivations, Ideology, and Policy”

- Hewlett Foundation Conference on Lobbying and Campaign Finance, October 2014, Palo Alto, CA

“Ideological Donors, Contribution Limits, and the Polarization of State Legislatures”

- Bipartisan Policy Center Meeting on Party Polarization and Campaign Finance, September 2014, Washington, DC

“Representing the Preferences of Donors, Partisans, and Voters in the U.S. Senate”

- Yale Center for the Study of American Politics Conference, May 2014, New Haven, CT

CONFERENCE  
PRESENTATIONS

Washington D.C. Political Economy Conference (PECO):

- 2017 discussant

American Political Science Association (APSA) Annual Meeting:

- 2014 participant and discussant, 2015 participant, 2016 participant, 2017 participant, 2018 participant, 2019 participant

Midwest Political Science Association (MPSA) Annual Meeting:

- 2015 participant and discussant, 2016 participant and discussant, 2018 participant, 2019 participant, 2020 (accepted, but not presented due to COVID-19)

Southern Political Science Association (SPSA) Annual Meeting:

- 2015 participant and discussant, 2016 participant and discussant, 2017 participant, 2020 (accepted, but not presented due to earthquake)

TEACHING  
EXPERIENCE

Poli 315: Congress and the Legislative Process

- Fall 2014, Winter 2015, Fall 2015, Winter 2016, Summer 2017, Fall 2018, Spring 2019

Poli 328: Quantitative Analysis

- Winter 2017, Fall 2017, Fall 2019, Winter 2020, Fall 2020, Winter 2021

Poli 410: Undergraduate Research Seminar in American Politics

- Fall 2014, Winter 2015, Fall 2015, Winter 2016, Summer 2017, Fall 2018

AWARDS AND  
GRANTS

2021 BYU Social Science College Research Grant, \$6,500

2020 BYU Social Science College Young Scholar Award

2019 BYU Mentored Environment Grant (MEG), Ideology in America Project, \$35,000

2017 BYU Political Science Teacher of the Year Award

2017 BYU Mentored Environment Grant (MEG), Funding American Democracy Project, \$20,000

2016 BYU Political Science Department, Political Ideology and President Trump (with Jeremy Pope), \$7,500

2016 BYU Office of Research and Creative Activities (ORCA) Student Mentored Grant x 3

- Hayden Galloway, Jennica Peterson, Rebecca Shuel

2015 BYU Office of Research and Creative Activities (ORCA) Student Mentored Grant x 3

- Michael-Sean Covey, Hayden Galloway, Sean Stephenson

2015 BYU Student Experiential Learning Grant, American Founding Comparative Constitutions Project (with Jeremy Pope), \$9,000

2015 BYU Social Science College Research Grant, \$5,000

2014 BYU Political Science Department, 2014 Washington DC Mayoral Pre-Election Poll (with Quin Monson and Kelly Patterson), \$3,000

2014 BYU Social Science College Award, 2014 Washington DC Mayoral Pre-Election Poll (with Quin Monson and Kelly Patterson), \$3,000

2014 BYU Center for the Study of Elections and Democracy, 2014 Washington DC Mayoral Pre-Election Poll (with Quin Monson and Kelly Patterson), \$2,000

2012 Princeton Center for the Study of Democratic Politics Dissertation Improvement Grant, \$5,000

2011 Princeton Mamdouha S. Bobst Center for Peace and Justice Dissertation Research Grant, \$5,000

2011 Princeton Political Economy Research Grant, \$1,500

ADDITIONAL  
TRAINING

EITM 2012 at Princeton University - Participant and Graduate Student Coordinator

COMPUTER  
SKILLS

Statistical Programs: R, Stata, SPSS, parallel computing

Updated March 8, 2021



I, Michael Barber, am being compensated for my time in preparing this report at an hourly rate of \$400/hour. My compensation is in no way contingent on the conclusions reached as a result of my analysis.

A handwritten signature in black ink, appearing to read "Michael Barber". The signature is cursive and somewhat stylized, with the first name "Michael" and last name "Barber" clearly distinguishable.

Michael Barber

March 9, 2021

RECEIVED

2021 MAR 11 PM 4:49

DEBRA P. HACKETT CLERK  
U.S. DISTRICT COURT  
MIDDLE DISTRICT ALA

## Exhibit 6

**UNITED STATES DISTRICT COURT FOR THE  
MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION**

THE STATE OF ALABAMA; ROBERT ADERHOLT, Representative for Alabama's 4th Congressional District, in his official and individual capacities; WILLIAM GREEN; AND CAMARAN WILLIAMS,

Plaintiffs,

v.

UNITED STATES DEPARTMENT OF COMMERCE; GINA RAIMONDO, in her official capacity as Secretary of Commerce; UNITED STATES BUREAU OF THE CENSUS, an agency within the United States Department of Commerce; and RON JARMIN, in his official capacity as Acting Director of the U.S. Census Bureau,

Defendants.

3:21-cv-211-RAH

CASE NO. \_\_\_\_\_

**DECLARATION OF THOMAS BRYAN**

THOMAS BRYAN pursuant to 28 U.S.C. § 1746, Federal Rule of Civil Procedure 26(a)(2)(B), and Rules 702 and 703 of the Federal Rules of Evidence, declares as follows:

1. I am 51 years old and competent to make this declaration.
2. I am an applied demographic, analytic, and research professional. I am the founder and principal of Bryan GeoDemographics, a demographic and analytic consultant firm

to meet the expanding demand for advanced analytic expertise in applied demographic research and analysis.

3. I have a Master's of Science in Management and Information Systems from George Washington University and a Master's of Science in Urban Studies with an emphasis in Demography and Statistics from Portland State University.

4. I have over 19 years of experience conducting complex demographic and analytical analyses, especially in the application of census.

5. My research and work focus includes:

- a. Redistricting and Voting Rights Act analysis;
- b. The application of U.S Census Bureau data;
- c. Large-scale multi-mode consumer survey research, design, and execution;
- d. Applied demographic techniques;
- e. Advanced analytics;
- f. Geographic Information Systems (GIS); and
- g. U.S. Government, Census, and other primary and secondary survey research data.

6. Plaintiffs requested that I assess the impact of the U.S. Census Bureau's approach to ensuring respondent privacy and Title XIII compliance by using a disclosure avoidance system involving differential privacy.

7. I am being compensated \$300 an hour for my time in connection with this matter. I am not being compensated for any specific opinion.

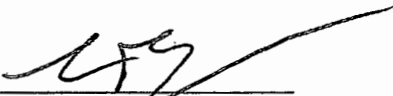
8. Attached and incorporated by reference to this declaration is my expert report in this matter. The report is attached hereto as Appendix 1.

9. Attached and incorporated by reference to this declaration is a copy of my curriculum vitae which lists, among other things, my qualifications, a list of all publications published over at least the last ten years, and a list of all cases over at least the past four years in which I

testified as an expert at trial or by deposition. My curriculum vitae is attached hereto as Appendix 2.

10. I declare under penalty of perjury that the foregoing, including any appendices, are true and correct according to the best of my knowledge, information, and belief.

Dated: March 9, 2021

  
\_\_\_\_\_  
Thomas Bryan

# Appendix 1

# **Census 2020**

## **Differential Privacy Analysis**

### **Alabama Case Study**

## **Census 2020 Differential Privacy Analysis Alabama Case Study**

- 1. Project Statement P3**
- 2. Alabama Demographics P4**
- 3. Differential Privacy Data, Analytic Approach and Findings P5**
  - a) PPMF Data Releases
  - b) Analytic Approach
  - c) Analytic Findings
    - 1) Case Studies: census block analysis
    - 2) Case Studies: non-voting age (NVA) children
    - 3) Analysis of impact on the 116<sup>th</sup> US Congressional Districts
    - 4) Analysis of impact on the Alabama State Legislative Districts
    - 5) Summary Statistics and Analysis at different levels of geography
- 4. Summary and Conclusions P41**
- 5. Appendices P42**
  - Appendix 1 Differential Privacy Data
  - Appendix 2 Terms
  - Appendix 3: 2010 – 2019 Estimated Total Population Changes in Alabama
  - Appendix 4: 2010 – 2019 Estimated Black African American Population Changes in Alabama
  - Appendix 5: 2010 – 2019 Estimated Hispanic Population Changes in Alabama
  - Appendix 6 Census
    - a) What is the Census?
    - b) Census Accuracy and Adjustments
    - c) Census Bureau Privacy, Confidentiality and Title 13 Privacy
    - d) Uses of the Census
  - Appendix 7 Differential Privacy
    - a) What is Differential Privacy?
    - b) How is Differential Privacy being proposed to be used in the 2020 Census?
    - c) Differential Privacy and the Census: Existing concerns from the user community



### Section 1 Project Statement

The purpose of the project is to assess the impact of US Census Bureau's proposed approach of ensuring respondent privacy and Title XIII compliance in the 2020 Census by using a Disclosure Avoidance System (DAS) involving Differential Privacy (DP). Conceptually, the Census Bureau is attempting to leverage DP to strike a balance between data quality (which would be reporting data as they were collected) and respondent privacy (which would be adjusting or "perturbing" data so dramatically that there would be no chance of anyone identifying a census respondent – which would also result in data that are practically no longer "quality" or of any use.).

The application of DP is a brand new approach for the Census Bureau and is different from all prior Census initiatives to comply with Title XIII. As the Census Bureau has been trying to develop the application of DP to their data, they have released a series of what they call data "demonstration products" to the public, including outside analysts and stakeholders, so they can determine for their purposes the impact DP would have on Census data. These demonstration products generally contain:

- the most common, basic demographic and housing variables;
- different levels of geography;
- data as they were originally reported in the SF (Summary Files) in 2010, which reported actual census data with small privacy protection modifications ; and
- trial data as they have been by adjusted (perturbed) DP.

This project seeks to determine the impact of DP on 2010 Summary file (SF) data for Alabama. We assess "spine" geography, which are standard census geographies such as counties and blocks, as well as "off-spine" geographies, which are political or administrative levels of geography such as cities, school districts, state legislative and senate districts, and congressional districts. This assessment is not only important for practical reasons, but it also enables us to uncover the unknown and oftentimes severe consequences. We believe these specific geographies are representative of the dataset as a whole and will enable us to reach reliable and valid conclusions about the data.

Ruggles et al. (2019: 406) argue that DP goes far beyond what is necessary to keep data safe under census law and precedent, and because it focuses on concealing individual characteristics instead of respondent identities, DP is a blunt and inefficient instrument for disclosure control. They go on to note that because the core metric of DP does not measure the risk of identity disclosure, it cannot assess disclosure risk as defined under census law, making it untenable for optimizing the privacy/usability trade-off.

If DP is implemented, it will affect almost all users of census data, from legislatures relying on the data to design Congressional and other districts to comply with the law, to demographics vendors who supply clients with zip code level characteristics so businesses can make better decisions. Other end users, such as health district administrators who need the data to track health issues such as COVID-19, and businesses that use small area data such as zip codes, blocks, and block groups to improve marketing, stand to be dramatically impacted. Many government agencies also depend on accurate small area census data to make programs run efficiently and effectively, and the biggest impact of DP will be in small areas. The data in small areas are typically used both directly where the small area is the unit of analysis and aggregated into higher levels of geography by these users.

The outcome of the project is a statement on the impact of the usability of 2020 Census data if it is subjected to DP, and alternatives available to the Census Bureau to protect privacy in the absence of DP. We use the most recent data available from the Census Bureau for each analysis.

Our study of the Alabama differential privacy census data leads us to conclude that it is a statistical adjustment of actual census data that make the data essentially unusable and unreliable at geographies below the statewide level for redistricting and other purposes.

### **Section 2: Alabama Demographics**

To better understand the implication of inaccuracies induced by the application of DP, we must first understand the demographics of Alabama itself. Since the last Decennial Census in 2010, the estimated size of the population in Alabama has not changed significantly: up only about +2% from 4,785,298 to 4,903,185 in 2019 (See Appendix 3). The demographic *complexion* of the state has changed dramatically though, according to estimates.

In this analysis, we examine changes in demographic estimates of Alabama from 2010 to 2019 by:

- Age (18+ / VAP and under 18 / NVA)
- Race (Black / African American and Hispanic)
- CVAP (citizen voting age population by race)

In assessing analysis of early versions of the DP datasets, one of the most significant observations is that the age structure at different geographies is materially changed. An accurate read of the size and changes of the population by age and race/ethnicity is critical for a wide variety of applications such as estimates and forecasts, strategic and infrastructure planning, and funding allocations.

In assessing changes in the total population of Alabama, there has been a significant change in citizenship from 2010-2019. The estimated number of NVA (children under 18) male foreign born dropped dramatically, while the estimated number of VAP who are naturalized grew significantly. This suggests a naturalization process for those males who aged in place. The estimated number of both NVA (under 18) and VAP female foreign born *grew* however, suggesting that this population may have immigrated and naturalized – rather than aging in place. This has profound implications for voting rights and redistricting in the state.

The estimated Black, non-Hispanic population growth outpaced total population growth. While the estimated total NVA (under 18) male foreign born dropped overall, it actually increased among Black, African Americans NVA. The estimated growth in total NVA (under 18) female foreign born was driven by Black, African American women (See Appendix 4).

The estimated Hispanic population growth outpaced all other demographic groups. Their numbers of foreign born dropped, while the numbers of those native and naturalized grew dramatically. Of particular note, the estimated total Hispanic CVAP in the state has nearly doubled. Where that growth took place will have profound implications for political representation and voting rights (See Appendix 5).

### **Section 3: Differential Privacy Data, Analytic Approach and Findings**

Two important related issues to consider in regard to applying this new technology called DP to the 2020 census are:

- 1) the level of testing it has undergone; and
- 2) the experience of Bureau staff with it.

Could applying DP to the 2020 Census be premature? As of the writing of this report at the beginning of March 2021, the latest view the public end users have of the data is a dataset from November 2020, which shows that the data are untenable and are fraught with contradictions, inconsistencies, and demographic impossibilities. DP has been in development at the Census Bureau for many years, and we are currently in the time frame we would be preparing for the release of the data under statutory timetables. And the Census Bureau has not yet produced a data product that is even remotely usable by the end user community – including state and local governments for the purpose of redistricting<sup>1</sup>.

---

<sup>1</sup> This is not the first time the Census Bureau has tried to push a new technology late into the process. The last case was when the Bureau attempted to automate the field data collection in the 2010 census in a program called “Field Data Collection Automation” (FDCA) headed by Deputy Director, Preston Jay Waite. The FDCA program was implemented shortly after the 2000 census was completed (Waite, 2003; Waite and Reist, 2005). Even though the FDCA program started well before the 2010 census, it turned into a debacle (Calleam Consulting, 2012), which resulted in the “early retirement” of Deputy Director Waite in 2008, two years before the 2010 census (PAA Affairs, Summer 2008., p. 6).

According to a report by the Congressional Research Service (Williams, 2012), the center piece of the FDCA program was the development of highly specialized handheld positioning software. Testing eventually revealed significant flaws in the handhelds, such as slow operation, memory problems, and a tendency to lock up when users entered large quantities of data. On April 3, 2008, in congressional testimony, then-Bureau Director Steve Murdock acknowledged that the Bureau had abandoned the plan to use the handhelds for Non Response Follow UP (NRFU) and instead would resort to the traditional paper-based approach and would rely on the handhelds only for address canvassing. The change required the Bureau to hire and train more NRFU staff, at significant increased expense. The GAO testified to Congress on June 11, 2008, that the Bureau had re-estimated the total life-cycle cost of the 2010 Census at between \$13.7 billion and \$14.5 billion, instead of the previously estimated \$11.5 billion. A 2009 House Committee on Appropriations report raised the estimate to \$14.7 billion. As the 2010 FDCA debacle demonstrated, the attempt to use an immature technology ended up not only causing problems that challenged the Census Bureau, but also leading to a significant increase in the cost of the 2010 census and the early retirement of Deputy Director Waite.

As the 2010 FDCA case suggests, there is room for concern with regard to applying a new method such as DP to the 2020 Census. Tests have been done in a shorter time frame than the 2010 FDCA tests, which means that it has not been as extensively vetted and Bureau staff have less experience with it than they had with the 2010 FDCA program. The fact that it took the Census Bureau nearly 20 years to get automated data collection right is a lesson to keep in mind when considering using DP for the 2020 census.

### **Part 3a: PPMF Data Releases**

In an effort to engage stakeholders, the Census Bureau released a series of demonstration products which showed how the application of differential privacy would have changed the 2010 Census data, had it been used. The National Historical Geographic Information System (NHGIS) linked these DP data with 2010 Census Summary File data to facilitate analysis and comparisons by end user groups. There have been four releases to date:

- **Census Demonstration Products v1 October 2019**  
In October of 2019, the Census Bureau released what they called a Demonstration Product, which consisted of the 2010 Census data with differential privacy applied. The IPUMS NHGIS team added the original census SF data to the differentially privatized data to make it easier for users to compare. In summarizing a review of this file by users, John Abowd (2020), the Census Bureau's architect of DP, wrote that much of the feedback identified areas where the disclosure avoidance system still needed to be improved.
- **Census Demonstration Products v2 May 2020**  
Another file was released by the Census Bureau in May of 2020, called a Privacy Protected Micro-Data File (PPMF). Concurrent with other concerns end users had about the Census Bureau's transparency and access to the files, many data users did not have the computational capacity to use such large files. It contained about 308 million records from the 2010 Census with DP applied. The IPUMS NHGIS team converted the PPMF to tables and added the same tables from the 2010 Census (SF) summary file. This is the last file released that has any age breakdown (in five year intervals) instead of 18 and over (VAP) and under age 18 (NVA).
- **Census Demonstration Products v3 September 2020**  
In September of 2020, the Census Bureau released another PPMF but only with data on the only age breaks on were the population over 18, and the population age 17 and under. This was related to the PL- 94-171 (redistricting file). After the file was issued, an error was identified<sup>2</sup> so the Census Bureau file had to re-process and re-release the data. It should be noted that while errors can and do occur in data releases, the fact that an external user found one suggests that the quantity and impact of changes introduced by the Census Bureau may make it impossible for them to be sure of the quality of all of the data they are releasing.
- **Census Demonstration Products v4 November 2020**  
Prior to the last data release from the Census Bureau, they reported, "Over the past several months, the Census Bureau has been making a number of improvements to the 2020 Census Disclosure Avoidance System (DAS) to address the concerns raised by the data user community at the December 2019 Committee on National Statistics workshop. Throughout this process, we have received numerous requests for additional tools to help evaluate this ongoing progress."<sup>3</sup>

---

<sup>2</sup> <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20200917/2020-09-17-erratum.pdf>

<sup>3</sup> <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20201116/2020-11-16-ppmf-factsheet.pdf>

Subsequently, the September 2020 files with the coding fixed was re-released in November 2020. This is the most recent data file available from the Census Bureau.

The Census Bureau has recently announced that they will produce one more demonstration product for users to evaluate by April 30, 2021 - and will make a final decision about how DP will be implemented in the redistricting data by early May 2021. Importantly, this leaves end users virtually no time for evaluation and no participation in the decision-making process. In other words, if the most recent data produced by DP is not adequate after several attempts, there is little time or opportunity left to make any other changes.

### **Section 3b: Analytic Approach**

The U.S. Census Bureau (2020a) has suggested several measures of accuracy that could be used to evaluate the data based on the application of DP to 2010 Census data. The Mean Absolute Error (Mean Absolute Numerical Error to distinguish it from the Mean Absolute Percent Error) and the Mean Absolute Percent Error are important summary measures. An absolute error reflects the magnitude of the error regardless of direction. This approach is used to make sure positive errors and negative errors do not cancel each other out and make it appear as if there are no errors. A geographic unit with an absolute error of 10 percent or more could be 10 percent too high or 10 percent too low. I focus on percent error because it reflects the size of the error relative to the size of the population. An error of a given magnitude (say 1,000 people) may be trivial in large places but very significant in smaller places.

Smaller geographic areas in terms of population size tend to have higher levels of error injected by DP. This is important because the census is designed to produce data for a lot of small geographic units. The vast majority of data produced in the U.S. Census are for small areas, and these small areas are where DP is designed to inject the most error percentage-wise. These errors are likely to cause problems in many use cases, such as the amount of state and federal funds received by school districts. For a small school district to get 10 percent less federal or state money than it would be entitled to with accurate census data will cause serious problems. It will be difficult for child advocates to support the use of DP in the 2020 Census if it produces significant errors like those identified in this paper.

The number and percent of large errors or outliers are the most important measures of accuracy. These extreme errors will be the biggest practical problem caused by DP. The fact that the biggest errors (percentage-wise) happen in smaller places is likely to generate concerns in many places across the state of Alabama. We seek to identify and capture these errors by leveraging a statistical method known as Loss Functions (Hough and Swanson, 2006).

In order to determine how much distortion is being inserted in the data by the application of differential privacy, we compare the 2010 Census data before and after the application of differential privacy. This allows us to measure the size of the inaccuracies caused by differential privacy. We perform a forensic examination of the original 2010 Census SF data and the 2010 DP data provided by the IPUMS NHGIS Privacy-Protected Demonstration Data (PPDD) for different demographic groups at different levels of geography<sup>4</sup>.

---

<sup>4</sup> <https://www.nhgis.org/privacy-protected-demonstration-data>.

Our analysis proceeds in five broad areas:

- 1) Case Studies: Block Analysis
- 2) Case Studies: underage non-voting age (NVA)
- 3) Impact on the 116<sup>th</sup> US Congressional District
- 4) Impact on the Alabama State Legislative Districts
- 5) Summary statistics and analysis at different levels of geography

A recent report commissioned by the Census Bureau concludes:

"To gain confidence around potential differential count of the population the Census Bureau should make use of its data science resources and summarize the assessments of data quality across various geographies and for relevant demographic groups. The report provided here responds to the recommendation for closer examination of geographies and demographic groups."<sup>5</sup>

Two different types of geographies are examined: "spine" which are the core census statistical geographies such as counties, tracts, and blocks, and "off-spine" which are governmental or administrative geographies such as school districts and legislative districts. The "spine" geography, particularly blocks, are important because they offer the greatest geographic granularity and are the geographies DP is actually being applied to. "Off-spine" geographies are also critically important because conceptually they could capture the best or worst pieces of statistical geography and aggregate and magnify their errors.

The levels of geography we processed for this analysis are:

- o census blocks
- o counties
- o unified school districts (USDs)
- o lower house districts (SLDLs)
- o upper house districts (SLDUs)
- o Congressional Districts (CDs)
- o Cities (Incorporated places and Census Designated Places / CDPs)

An additional important concept that warrants understanding is "invariants". On November 24th, the Census Bureau's Data Stewardship Executive Policy Committee (DSEP) finalized the list of "invariants" for the first set of 2020 Census data products. Invariants are statistics that are published without DP. Per the decision, the following statistics will be invariant at these levels of geography and higher:

- Total population (at the state and state-equivalents level)
- Total housing units (at the census block level)
- Number of group quarters facilities by type (*NOT* actual GQ pop, at the census block level)

Aside from these invariants, every other population (such as by age or race / ethnicity) can be impacted by DP at any level of geography.<sup>6</sup>

---

<sup>5</sup> JONAS (2021, pages 8) Letter Report to Christa D. Jones and Deborah M. Stempowski, U.S. Census Bureau dated February 8, 2021, <https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-census-data-quality-processes.pdf>

<sup>6</sup> <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>

### **Part 3c. Analytic Findings**

#### **Case Inventory**

Moving now to individual cases of examination. There are two general types of case studies are shown below here. First, the implications for significant changes to census blocks for a variety of populations. Second, we focus on the implications for one specific population, NVA children.

#### **3c1 Census Block Analysis**

- Case 1: Children without Adults: Differential Privacy turned 5 blocks into 13,842.
- Case 2: DP turned 30,338 blocks with one or more VAP into blocks with zero VAP.
- Case 3: DP turned even more blocks with one or more VAP into blocks with zero VAP by race.
- Case 4: Differential Privacy turned 46,730 Blocks with one or more people of non-voting age into blocks with zero people of non-voting age.
- Case 5: Blocks with Extreme Differences between NVA and VAP.
- Case 6: Plaintiff Green's Situation
- Case 7: Household Population and Occupied Housing Unit Inconsistencies.

#### **3c2 NVA Children Cases**

- Case 8: Implications of DP at Unified School Districts for NVA Children.
- Case 9: Implications of DP at Census Tracts and Counties on Preschoolers.
- Case 10. Unrealistic sex ratios for young children.

#### **3c1 Case Studies: Census Block Analysis**

In this analysis, 137,081 census blocks with population were extracted from the total of 252,266 blocks in Alabama. This extraction/excluded blocks in which zero people were reported in both the 2010 Census and the DP file built from the 2010 Census (115,185 blocks) leaving 137,081 blocks populated in one, the other or both DP and SF data.<sup>7</sup> This breaks out as:

- a) 118,495 blocks have population in both the SF and DP file
- b) 16,944 blocks have population in the SF file, but not the DP file
- c) 1,642 blocks have population in the DP file, but not the SF file

In order to grasp the severity of the impact of the DP process on the data, we examined how many pieces of Alabama block geography were populated in the 2010 SF file, and how many blocks had populations that were perturbed by DP. In Column 1 below, we see that there were 135,439 total populated blocks (bullet a) above with 118,495 blocks + bullet b) above with 16,944 blocks – not including bullet c) with 1,642 blocks that were previously unpopulated in the SF file – but were populated by DP).

---

<sup>7</sup> The 2010 Census saw a substantial increase in the number of blocks from the 2000 Census. After the 2010 Census data users commented that some of the new census blocks were not useful, particularly very small blocks. Many unnecessary small blocks were formed by incorrect roads that had not been deleted from the Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Database as well as misclassified highway ramps, traffic circles, cul-de-sacs, alleys, and minor unnamed roads. In addition, small water bodies and overly detailed or incorrect water features contributed to the increase in unnecessary small blocks. These types of blocks make up the majority of zero population blocks in the SF file in Alabama.

**Table 3c1: Number and Percent of Demographic Groups Whose Populated Blocks Changed**

<b>Population</b>	<b>Column 1 # of Blocks 2010 SF &gt; 0 Pop</b>	<b>Column 2 # of Blocks Changed SF&gt;DP</b>	<b>Column 3 % of Blocks Changed</b>
Total	135,439	127,809	94%
Total Hispanic	26,952	26,387	98%
Total White, non-Hispanic	116,998	112,180	96%
Total Black, non-Hispanic	59,878	57,950	97%
Total, Other <sup>8</sup> non-Hispanic	34,457	33,700	98%
Voting Age Population (VAP)	135,434	129,837	96%
VAP Hispanic	24,933	24,541	98%
VAP White, non-Hispanic	116,878	113,055	97%
VAP Black, non-Hispanic	59,393	57,871	97%
VAP Other non-Hispanic	30,226	29,665	98%
Non-Voting Age Pop (NVA)	103,945	100,905	97%
NVA Hispanic	16,115	15,811	98%
NVA White, non-Hispanic	81,057	79,188	98%
NVA Black, non-Hispanic	42,381	41,521	98%
NVA Other non-Hispanic	17,494	17,234	99%

There were 252,266 blocks for the 2010 Census. As context for the examples below – many blocks in Alabama had zero population to begin with in SF, before DP was introduced and zeroed out many more. Column 1 shows the number of blocks populated in 2010. Column 2 shows how many populated blocks were changed by DP, and Column 3 shows the percent of populated blocks that were changed by DP.

**Total reading example:** There are 135,439 total populated blocks in the SF file. Of these, 127,809 have their population changed – either to zero or some other number by DP. This represents 94% of all VAP populated blocks.

**VAP reading example:** There are 135,434 VAP populated blocks in the SF file. Of these, 129,837 have their VAP population changed – either to zero or some other number by DP (as shown in Case 2 below, 30,338 of these blocks had VAP population that was zeroed out by DP). This represents 96% of all VAP populated blocks. Note: the difference between the 135,439 total populated blocks and the 135,434 blocks here are the 5 blocks occupied by NVA children alone in the SF file. As we will see below in Case 1: the block count where there are children but no adults swells from 5 in the SF file to 13,842 in the DP file.

**NVA reading example:** There are 103,945 NVA populated blocks in the SF file. Of these, 100,905 have their NVA children population changed – either to zero or some other number by DP (as shown in Case 4 below, 46,730 of these blocks had NVA population that was zeroed out by DP). This represents 97% of all VAP populated blocks.

<sup>8</sup> Includes Asian, Native Hawaiian and Pacific Islander, American Indian and Alaskan Native, reported “Other” and 2+ multi-race – all non-Hispanic.



**Table 3c2: Changes Between SF and DP Populations**

The following table highlights the differences between the DP and SF measurement of VAP and NVA populations. For example, in the second to last row, I show that there are 46,730 blocks in the SF file that have NVA children but have zero children in the DP data. We explore this finding in Case 4. This suggests that a frequent outcome of the method is not that DP swaps existing data from another geography to ensure it retains some of its original fidelity, but DP simply deletes data in wholesale fashion if there's a concern.

VAP DP	NVA DP	VAP SF	NVA SF	Blocks	Case
		0	>0	5	Case 1
0	>0			13,842	Case 1
0		>0		30,338	Case 2
>0		0		1,199	Case 2
	0		>0	46,730	Case 4
	>0		0	5	Case 4

**Case 1: Children without Adults: Differential Privacy turned 5 blocks into 13,842.**

The 2010 Census reported in the SF file that there were five blocks in which 1 or more children (under age 18) were listed, but no adults (18 years and over). Of these five blocks, the first had one child, the second, 11 children; the third, 22 children; the fourth, 23 children; and the fifth block, 74 children. It is likely that most if not all of these five blocks have facilities where children reside in the presence of adults who themselves live elsewhere. By comparison, in the DP file there are 13,842 blocks in which there are no VAP adults, and there is at least one NVA child living there.

Out of 137,081 populated blocks in Alabama, it is highly believable that there are five blocks in which only children reside. Juvenile group quarters for example. However, the Differential Privacy Algorithm produced 13,842 such blocks, a *highly* unbelievable number. Ten percent of the blocks examined have children residing alone according to DP. In these blocks there are now 31 blocks in which 50 or more children reside alone, four of which have more than 70 children residing alone. And in between, there are 13,810 blocks with between two and 49 children residing without adults. In total, where the 2010 Census had 131 children residing in five blocks without adults, DP produces over 141,817 children residing in 13,842 blocks without adults.

**Case 2: DP turned 30,338 blocks with one or more VAP into blocks with zero VAP.**

This analysis uses the same 137,081 DP- or SF- populated census blocks in the preceding example.

- In comparing the voting age populations reported by the 2010 Census and the DP file, it was found that there are 30,338 blocks in which DP reported zero VAP while the SF reported one or more VAP in these same blocks. These blocks were populated with 165,744 VAP.
- At the same time, DP turned 1,199 blocks in which SF reported zero persons of voting age into blocks with >0 persons of voting age.

**Case 3: DP turned even more blocks with one or more VAP into blocks with zero VAP by race.**

Again using the same 137,081 populated census blocks found in the preceding cases, it was found that there were:

- 19,666 blocks in which DP reported zero Hispanic persons of voting age while the 2010 Census reported one or more Hispanic persons of voting age in these same blocks;
- 38,568 blocks in which DP reported zero White non-Hispanic persons of voting age while the 2010 Census reported that one or more White non-Hispanic persons of voting age were in these same blocks; and
- 38,010 blocks in which DP reported zero Black Non-Hispanic persons of voting age, while the 2010 Census reported one or more Black non-Hispanic persons in the same blocks.

Looking in the opposite direction, there were:

- 7,384 blocks in which the 2010 Census reported 1 or more Hispanic persons of voting age while DP reported zero Hispanic persons of voting age in these same blocks;
- 4,202 blocks in which the 2010 Census reported 1 or more White non-Hispanic persons of Voting age while DP reported these same blocks to have zero White non-Hispanic persons of voting age; and
- 8,073 blocks in which the 2010 Census reported 1 or more Black non-Hispanic persons of voting age while DP reported zero Black non-Hispanic persons of voting age in these same blocks.

**Case 4: Differential Privacy turned 46,730 Blocks with one or more people of non-voting age into blocks with zero people of non-voting age.**

This analysis uses the same 137,081 census blocks found in the preceding example.

- In comparing the voting age populations reported by the 2010 Census and the DP file, it was found that there are 46,730 blocks in which DP reported zero people of voting age while the SF reported one or more persons of voting age in these same blocks.
- At the same time, DP turned 5 blocks in which SF reported zero non-voting age population into blocks with >0 persons of non-voting age.

**Case 5: Blocks with Extreme Differences between NVA and VAP**

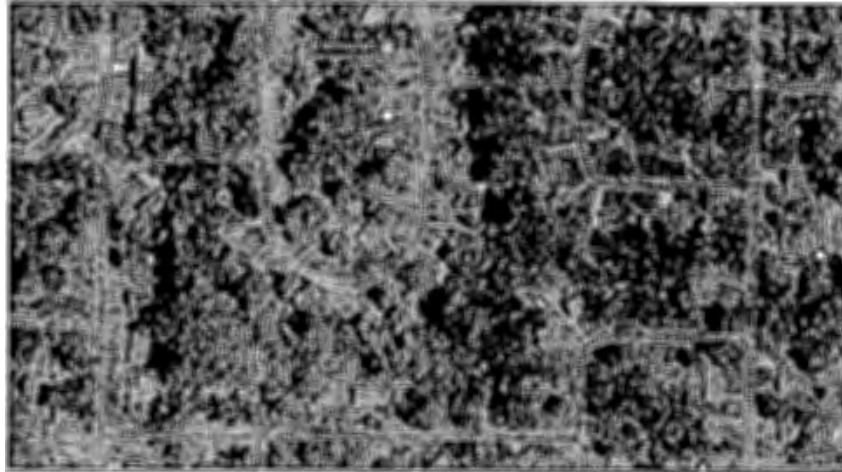
Of the 13,842 blocks containing over 141,000 NVA children and no VAP, four blocks stand out with more than seventy children each in them:

<u>Block</u>	<u>NVA Pop. DP</u>
010479570001345	76
010730019021024	77
010730118032035	82
010970024001008	72

Block 010479570001345 is a large, poorly defined heavily rural area west of Carlowville, AL. There is a juvenile GQ facility there. This is plausible.

Block 010730019021024 is the location of the Gateway-Rushton School, a juvenile GQ. There were 25 NVA and 11 VAP there in the SF data. Now there are no adults and 77 residents. This change is not plausible.

Block 010730118032035 is a tree lined single family neighborhood on the north side of Birmingham, where it is simply implausible that there are no adults.



Block 010970024001008 is a tree lined single family neighborhood in Mobile south of US 90 where again it is simply implausible that no adults live here.



### **Case 6: Plaintiff Green's Situation**

Plaintiff William Green resides in a block that reflects the all-too-common changes DP inflicts on previously accurately reported census data. Plaintiff Green resides in census block 011010031003014. This block was reported in the 2010 SF file to have 22 residents, 21 of whom were Black, non-Hispanic. 15 of these residents were VAP, and 14 of these were Black, non-Hispanic. 7 of these residents were NVA children, all of whom were Black, non-Hispanic.

With DP introduced – the characteristics of Plaintiff Green's block changes dramatically. The block is increased to have 27 residents, but only 9 of whom are now Black, non-Hispanic with the remaining majority of 18 being Hispanic or other, non-Hispanic. 21 of these residents are now VAP, but only 9 of whom are now Black, non-Hispanic with the remaining majority of 12 being Hispanic. All 7 of the Black, non-Hispanic children are removed and replaced with 5 Hispanic and 1 other, non-Hispanic child.

### **Case 7: Household Population and Occupied Housing Unit Inconsistencies**

In the 2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop,<sup>9</sup> Beth Jarosz (Population Reference Bureau) states that accuracy and internal consistency were key for planning: "Planners look at several mathematical identities in assessing local demographics, for instance: (1) population must equal household population plus group quarters population; (2) occupied units must be the sum of all housing units minus vacant units; and (3) average household size must be at least one, and household size multiplied by number of occupied units gives the population in a size category. None of these identities held for every jurisdiction in the PPMF."

In the SF data, there are no cases where there are more occupied housing units than household population. It is a demographic and logical impossibility. In the DP data, there are 22,404 cases where there are more occupied housing units than household population. Additionally, there are 15,288 blocks where there are occupied housing units, but no household population.

### **Case Studies: non-voting age (NVA) children**

In this section the implications of DP are examined for a specific population, namely children (population ages 0 to 17). It is likely that the implications for children are seen in other population groups, but children are the focus here for a couple of reasons. First, DP infused data are provided for Unified School Districts so the implications for a key public institution can be examined. Also, there are already a number of studies that examined the implication of DP for children and schools that can be built on. (O'Hare 2020, Nagle 2019, Sojourner 2019)

Total federal spending on children is currently \$325.4 billion (First Focus on Children, 2019) Children's Budget 2019. Within this budget, schools (public education institutions) represent the largest share with \$39 billion distributed by the U.S. Department of Education (Reamer, 2020).

---

<sup>9</sup> 5.3.2 Housing and Population Consistency Page 87

**Table 3c3**

Selected Federal Expenditure Programs Focused on Young Children Guided by Census-Related Data , FY2016 Distributions (2-28-2021)		
	Dollars FY 2016	
	U.S	Alabama
Head Start	\$8,648,933,810	\$138,342,659
Supplemental Nutrition Program for Women, Infants, and Children	\$6,383,830,000	\$110,726,000
Child Care Mandatory and Matching Funds	\$2,840,075,000	\$41,247,000
Child Care and Development Block Grant	\$2,612,564,000	\$50,468,000
Total	\$20,485,402,810	\$340,783,659
Source Counting for Dollars website <a href="https://gwipp.gwu.edu/sites/g/files/zaxdzs2181/f/downloads/Characteristics%20of%205%20Large%20Census-guided%20Programs.pdf">https://gwipp.gwu.edu/sites/g/files/zaxdzs2181/f/downloads/Characteristics%20of%205%20Large%20Census-guided%20Programs.pdf</a>		

The analysis in this section used data from the May 27, 2020 Census Bureau release because that is the most recent data from the Census Bureau that allows one to examine young children (ages 0 to 4) as well as school-aged children (ages 5 to 17).

The use of DP severely impacts the accuracy of data for young children (ages 0 to 4) and for school-aged children (ages 5 to 17) for school districts in Alabama. These age groups are the focus of the analysis shown later in this report because data on the preschool population (ages 0 to 4) are used to forecast future education needs and the needs for day care centers. Data for school-aged children (ages 5 to 17) are used to allocate state and federal funding to localities. Errors in the Census will impact the fairness of such distributions.

The impact of DP on preschoolers (ages 0 to 4) as well as school-aged (ages 5 to 17) are examined in the context of unified school districts in Alabama. Then the population 0 to 4 is examined in the context of other kinds of geographic units in Alabama.

Metrics for assessing the accuracy of census data for two age groups of children—age 0 to 4 (preschoolers) and ages 5 to 17 (school-age populations)—are assessed here by reporting empirical evidence about the likely level of errors injected into the Census data by DP for children based on the most recently available data from the Census Bureau.

#### **Case 8: Implications of DP on Unified School Districts and NVA Children**

The first focus is on the population ages 0 to 4. In their March 2020 release, the U.S. Census Bureau (2020a) provided data related to several “Use Cases” and the population ages 0 to 4 was one of those. Table 5d5b provides several accuracy measures for the population ages 0 to 4 for 134 unified school districts in Alabama (in the 2010 Census).

**Table 3c4**

Summary Statistics for Application of Differential Privacy for 134 Unified School Districts in Alabama (2-28-2021)		
	Ages 0 to 4	Ages 5 to 17
Mean Absolute Numerical Error	153	104
Mean Absolute Percent Error	9.8	2.8
Percent of School Districts with Errors of 10 percent or more	43	1
Percent of School Districts with Errors of 5 percent or more	65	19
Source: Analysis of Census Bureau data released in May 2020		
*Error is defined here as the difference between the data reported in the 2010 Census and the data after differential privacy was applied.		

Number with 10% or more error: 58 for ages 0-4 and 1 for ages 5-17  
 Number with 5% or more error: 87 for ages 0 to 4 and 25 for ages 5-17

Sometimes positive and negative errors cancel each other out, so it is important to look at absolute errors. Absolute errors reflect the magnitude of the error regardless of the direction (i.e., positive or negative). The mean absolute numerical error for the population ages 0 to 4 in Case Study Table 1 is 153. That means, on average, the number of children ages 0 to 4 produced by DP infused data was 153 children different than the 2010 Census count based on data from respondents. Since the average preschool class size is about 17 (Samuels, C.A. 2017), that means the average error of 153 children represent about nine classrooms.

The mean absolute percent error for ages 0 to 4 was 9.8 percent. This means on average across the unified school district of Alabama, there was nearly a 10 percent error in the number of young children ages 0 to 4. Perhaps the more important data in the table is the number of school districts that are likely to have large errors based on the application of DP. For NVA ages 0 to 4, 43 percent of the school districts displayed errors of 10 percent or more, and 65 percent experienced errors of 5 percent or more after DP was applied.

The 2010 Census SF data reported that Midfield City School District children ages 0 to 4 was 405, but after DP was applied to the data, it was 540. This is an increase of 135 children. The average class size for preschools in Alabama is about 18 children.<sup>10</sup> The error of 135 preschool children in Midfield City School Districts amounts to about 8 classrooms.

The next focus is NVA ages 5 to 17. The mean absolute numerical error for the population ages 5 to 17 in Table 1 is 104 children. That means, on average, the number of children ages 5 to 17 produced by DP infused data was 104 children different than the 2010 Census count based on respondents' input. The mean absolute percent error for ages 5 to 17 was 2.8 percent.

Perhaps the more important data in Table 1 is the number of school districts that are likely to have large error based on the application of DP. For ages 5 to 17, only 1 school district had an error of 10 percent or more, but 19 percent experienced errors of 5 percent or more after DP was applied.

There are several large changes after DP was applied to the census data that stand out. For example, the 2010 Census reported that Clarke County School District had 1,295 children ages 0 to 4, but after DP was

<sup>10</sup> <https://nieer.org/wp-content/uploads/2016/08/9.pdf>

applied, the number of children ages 0 to 4 was decreased to only 885. This is a reduction of 410 children, or 32 percent.

According to the National Center for Education Statistics, the average class size for public schools in Alabama is about 20 students. The error of 410 students for Clarke County School Districts amounts to about twenty classrooms. If 410 unexpected students show up in the Clarke County School Districts, that will lead to crowded classrooms. On the other hand, building and staffing 20 classrooms that are unneeded because of inaccurate census data would be problematic. That is why getting accurate data on the school-age population is so important.

The school district with the largest decrease in the number of children ages 0 to 4 was Mobile County School District. The 2010 Census reported 28,201 children ages 0 to 4, but after DP was applied to the reported census data, the figure was decreased to 27,358. This is a decrease of 843, or 3.0 percent. 843 children is the population size of one or two elementary schools.

The 2010 Census reported that Midfield City School District had 1,130 children between the ages of 5 and 17, but after DP was applied to the 2010 Census data, the number of children ages 5 to 17 was only 1,015. This amounts to a reduction of 115 children, or 10.2 percent.

The largest numerical decrease for school-aged children was seen in Madison City School District. The 2010 Census reported 9,548 children ages 5 to 17, but after DP was applied, the figure was changed to 8,774. This is a decrease of 776, or 9.0 percent.

Many other researchers have noted potential problems related to the infusion of differential privacy for data used by school systems. For example, in examining the impact of differential privacy on school-age population in Cambridge, MA, (Cook 2019 page 60) Cook noted that the 2010 DP added 800 5-17 year-olds, for the city compared to SF1, which is a number big enough to justify another elementary school.

The most recent data available from the U.S. Census Bureau regarding the likely impact of DP on 2020 Census data for children suggests that the level of error introduced will result in a high level of errors for many unified school districts in Alabama for both the pre-school population (ages 0 to 4) and the school-age population (ages 5 to 17).

In the previous section, the importance of the population ages 0 to 4 was discussed in the context of school systems. The number of young children in a community has several important ramifications. The number of young children is often used to forecast future school enrollment, which has implications for hiring staff, building facilities, and establishing curriculum. A number of federal programs designed to aid young children provide funding based on the number of young children in a community. Table 5d5a above shows data for four such programs. These federal programs provided \$341 million for young children in Alabama in Fiscal Year 2016 (above)

In addition to federal programs, many states and localities provide funds to help young children and their families. The number of preschool children in a community drives the need for childcare as well as things like playgrounds. Young children are one of the most vulnerable populations in Alabama. The poverty for children ages 0 to 4 in Alabama, based on the 2019 American Community Survey, was 24 percent compared to 20 percent for children ages 5 to 17, 15 percent for working age adults (ages 18 to 65), and 11 percent for seniors (ages 65 plus). In other words, the poverty rate for young children (ages 0 to 4) is more than twice that of seniors.

**Table 3c5**

Poverty Rate by Age in Alabama: 2019 (2-25-2021)	
	Percent in Poverty
Age 0 to 4 (Preschool Population)	24
Age 5 to 17 (School-Ages Population)	20
Age 18 to 64 (Working Age adults)	15
Ages 65 plus (Seniors)	11
Source: U.S. Census Bureau, American Community Survey, 2019 Table ID S 1701	

For Black and Hispanic young children, the poverty rates are even higher. Therefore, inaccurate data for young children can have important public policy implications and result in misappropriation of public funds and resources.

**Case 9: Implications of DP on Census Tracts and Counties for Preschoolers**

On average, the application of differential privacy changed the number of young children (ages 0 to 4) in the census tracts in Alabama by 41. That is roughly the size of two preschool classes. The average absolute percent error for the 1,174 tracts in Alabama for ages 0 to 4 was 19.3 percent. For agencies trying to decide the need for preschool facilities, an error of this magnitude could be very problematic. Building two extra preschool classrooms when they are not needed would be problematic. On the other hand, if the number of young children who show up for preschool is much larger than anticipated based on the data, it will result in overcrowded classrooms, inappropriate student/teacher ratios, and other complications.



**Table 3c6**

Summary Statistics for Application of Differential Privacy for Population Ages 0 to 4 for 1,174 Census Tracts in Alabama (2-28-2021)	
Mean Absolute Numerical Error*	41
Mean Absolute Percent Error*	19.1
Percent of Places with Errors of 10 percent or more (# =771)	66
Percent of Places with Errors of 5 percent or more (# = 962)	82
Source; Analysis of Census Bureau data released May 27, 2020	
* Error is defined here as the difference between the data as reported by the Census respondents and the data after the application of DP.	
Seven Census tracts were not included in the analysis because they had zero populaiton.	

There are some census tracts where the problem was particularly acute. In census tract 55.04, the 2010 Census reported 56 children ages 0 to 4 based on respondents' input, but after the application of differential privacy, the number was changed to 137. In census tract 7.02, the 2010 Census reported 105 children ages 0 to 4 based on respondents' input, but after the application of differential privacy, the number was changed to 188. There are many other examples like this. There were 33 census tracts where the number of children ages 0 to 4 reported after application of differential privacy was more than 100 children different than the number reported in the Census. It is not difficult to imagine how this misinformation could be problematic for someone trying to design services for preschoolers.

When differential privacy is applied to the data from the 2010 Census, the number of young children (ages 0-4) is changed significantly in many counties in Alabama. Of the 67 counties in Alabama, there were 16 counties where the number of young children was altered by more than 100 after differential privacy was applied.

Some of the distortions were extreme. For example, the 2010 Census reported there were 2,385 children ages 0 to 4 in Escambia County, but after differential privacy was applied, that number was changed to 2,044. The difference is 341. This amounts to roughly the number of young children there would be in 20 preschool classes. Another example is Clarke County. In Clarke County, the 2010 Census reported that there were 1,468 young children in Clarke County, but after differential privacy was applied that number was changed to 1,256. This amounts to a difference of 212, which is the equivalent of about 12 preschool classes.

**Case 10: Unrealistic Sex Ratios for Young Children**

The sex ratio (number of males divided by the number of females, times 100) is one the most fundamental demographic measures available. Especially for young children, the number of males should generally be nearly equal to the number of females. If the numbers of young males and females are substantially different, it begs an explanation. If there is not a reasonable explanation, it suggests the data are erroneous. If there are large numbers of geographic units where the number of males is much larger or much smaller than the number of females, it is highly improbable.

Sex ratios of young children in census tracts were examined. The table below shows that, before differential privacy was applied to the 2020 Census data, there were no census tracts where the number of males was more than 100 more or more than 100 less than the number of young females. On the other hand, the table below shows that there were 67 tracts where the number of males was more than 100 higher than the number of females, and 53 census tracts where the number of males was 100 less than females after DP had been applied to the 2010 Census data.

**Table 3c7**

Sex Ratios in 2010 Census Tracts in Alabama for Ages 0 to 4 With and Without Differential Privacy (2-28-2021)		
	Number of Tracts where there were at least 100 more males than females	Number of Tracts where there were at least 100 fewer males than females
2010 Census With Differential Privacy	69	53
2010 Census Without Differential Privacy	0	0
Source: Analysis of file released by the Census Bureau in May 2020		

Some of the situations are extreme. For example, after differential privacy had been applied to 2010 Census Data, Tract 27 had 57 males ages 0 to 4 and 14 females ages 0 to 4. Also, tract 401.05 had 73 males ages 0 to 4 and 28 females ages 0 to 4.

This indicates the application of differential privacy converted mostly reasonable statistics into a large number of statistics that were not reasonable.

**Analysis 3c3: Impact on the 116th US Congressional Districts**

The data tables provided by IPUMS for the 116<sup>th</sup> Congressional districts are for the data in the pre-2010 Congressional geography. So, we undertook the exercise of assigning IPUMS block data to Alabama's 116<sup>th</sup> Congressional districts and summarized to reflect Alabama's data post-redistricting.

In this analysis and subsequent analyses, I use thematic shading to highlight significant values and differences. Cells highlighted in green illustrate large values and positive differences. Cells highlighted in red illustrate small values and negative differences.

**Figure 3c3 DP Population, SF Population and Difference by 116<sup>th</sup> Congressional Districts<sup>11</sup>**

DP Data	Total Pop	Hispanic Pop	WNH Pop	BNH Pop	ONH Pop
1	682,747	19,114	449,222	188,251	26,160
2	682,791	24,486	437,298	200,530	20,477
3	682,844	17,959	474,706	170,881	19,298
4	682,820	39,219	579,848	45,689	18,064
5	682,820	32,630	506,301	114,680	29,209
6	682,688	33,215	537,604	92,201	19,668
7	683,026	19,180	219,557	432,227	12,062
<b>Grand Total</b>	<b>4,779,736</b>	<b>185,803</b>	<b>3,204,536</b>	<b>1,244,459</b>	<b>144,938</b>

SF Data	Total Pop	Hispanic Pop	WNH Pop	BNH Pop	ONH Pop
1	682,820	19,087	449,560	187,883	26,290
2	682,820	24,612	437,289	200,187	20,732
3	682,819	17,958	474,877	170,713	19,271
4	682,819	38,949	579,614	46,166	18,090
5	682,819	32,562	506,130	114,885	29,242
6	682,819	33,345	537,197	92,020	20,257
7	682,820	19,089	219,735	432,583	11,413
<b>Grand Total</b>	<b>4,779,736</b>	<b>185,602</b>	<b>3,204,402</b>	<b>1,244,437</b>	<b>145,295</b>

Difference	Total Pop	Hispanic Pop	WNH Pop	BNH Pop	ONH Pop
1	-73	27	-338	368	-130
2	-29	-126	9	343	-255
3	25	1	-171	168	27
4	1	270	234	-477	-26
5	1	68	171	-205	-33
6	-131	-130	407	181	-589
7	206	91	-178	-356	649
<b>Grand Total</b>	<b>0</b>	<b>201</b>	<b>134</b>	<b>22</b>	<b>-357</b>

These differences are larger than those that have resulted in at least one federal court striking down congressional redistricting plans in the past<sup>12</sup>.

<sup>11</sup> Hispanic Pop: population includes all races, WNH Pop: White, non-Hispanic population, BNH Pop: Black, non-Hispanic population, ONH: Other non-Hispanic population, including American Indian and Alaskan Native, Asian, Native Hawaiian and Pacific Islander, Two + and Other races

<sup>12</sup> Veith v. Pennsylvania 195 F. Supp. 2D 672 (M.D.Pa. 2002)

**Analysis 3c4: Impact on the Alabama State Legislative Districts**

The data tables provided by IPUMS for the 105 Alabama State Legislative districts are for the data in the pre-2010 Congressional geography. So, we undertook the exercise of assigning IPUMS block data to Alabama State Legislative districts and summarized to reflect Alabama's data post-redistricting.

**Table 3c4a: State Legislative Districts Difference between SF Pop and DP Pop by Race / Ethnicity**

District	Total Pop	Hispanic Pop	WNH Pop	BNH Pop	ONH Pop
1	-8	37	202	197	40
2	-9	-2	121	-78	-50
3	-47	-11	31	-3	-64
4	318	15	224	136	-57
5	-259	-67	-152	41	-81
6	-20	118	110	-82	-166
7	-95	92	-171	2	-18
8	0	-11	125	-251	137
9	-106	53	-17	11	-153
10	2	-30	-98	259	-129
11	126	134	63	-32	-39
12	46	41	-24	-2	31
13	-5	26	182	-73	-140
14	17	48	-48	30	-13
15	129	28	108	71	-78
16	25	14	-95	115	-9
17	1	5	-25	-84	105
18	20	49	20	-175	126
19	37	-50	54	-37	70
20	-19	-88	131	23	-85
21	-24	119	-224	10	71
22	137	54	109	30	-56
23	-47	-3	-38	-110	104
24	-72	29	-22	-94	15
25	-6	-53	-21	-162	230
26	-58	-42	-134	-97	215
27	-134	-147	95	-77	-5
28	161	4	312	-225	70
29	56	59	-120	148	-31
30	-184	-21	-170	144	-137
31	-40	27	-31	-137	39
32	-42	-26	105	-269	148
33	-62	17	-152	45	28
34	36	62	28	5	-59
35	68	-17	-62	177	-30
36	88	153	-66	51	-50
37	-11	-95	108	-20	-4
38	74	15	79	-39	19
39	27	-60	39	39	9
40	0	-4	76	-41	-31
41	-6	-8	27	-46	21
42	-132	50	-340	201	-43
43	-22	-122	117	27	-44
44	-36	-43	100	-121	28
45	-61	48	-5	-108	4
46	80	5	74	195	-194
47	32	-4	-24	19	41
48	-33	2	114	-102	-47
49	-239	49	-84	-105	99

Source: 20201116 PPMF

**Table 3c4a: State Legislative Districts Difference between SF Pop and DP Pop by Race / Ethnicity Cont.**

District	Total Pop	Hispanic Pop	WNH Pop	BNH Pop	ONH Pop
50	2	-62	27	79	-42
51	79	-136	137	146	-68
52	-25	16	29	-57	-13
53	13	-76	98	-136	127
54	55	-81	18	-4	122
55	68	-53	-175	144	152
56	-149	49	-102	-95	-1
57	8	45	51	-67	-21
58	1	95	350	204	52
59	-52	91	48	-215	24
60	-47	-17	40	-127	57
61	-21	-62	115	-85	11
62	-15	-126	54	11	46
63	59	129	85	-121	-34
64	-43	112	-173	-3	21
65	105	93	-156	216	-48
66	-43	-16	-136	97	12
67	-17	-34	-84	111	-10
68	-166	-5	241	399	-3
69	-11	-11	10	73	-83
70	-64	-32	-25	70	-77
71	1	-7	95	-48	-39
72	33	28	-17	71	-49
73	242	41	5	246	-50
74	8	70	12	19	-93
75	48	-37	-37	176	-54
76	65	-66	68	72	-9
77	13	-145	30	93	35
78	-126	-7	-84	-201	166
79	-88	64	-141	175	-186
80	-18	-80	30	-28	60
81	177	95	12	4	66
82	-16	120	8	-266	122
83	-18	-129	155	-9	-35
84	86	58	-56	82	2
85	-19	-15	36	-107	67
86	10	-45	-151	-193	13
87	-12	0	220	150	-82
88	46	69	17	-173	133
89	-91	55	20	-198	32
90	33	26	-19	51	-25
91	-27	-37	-49	156	-97
92	43	33	2	50	-42
93	123	-79	-6	279	-71
94	28	-213	270	-20	-9
95	-6	-114	-239	143	-24
96	109	-14	-149	222	50
97	-39	-53	156	339	197
98	-14	-83	-26	29	66
99	4	139	183	25	23
100	30	-34	-13	-43	120
101	-52	39	29	79	-199
102	-49	-27	36	58	-116
103	41	-65	-84	26	164
104	-88	8	31	-11	-116
105	83	10	115	118	-160
<b>Grand Total</b>	<b>0</b>	<b>201</b>	<b>134</b>	<b>22</b>	<b>-357</b>

Source: 20201116 PPMF

Table 3c4b: State Legislative Districts DP and SF VAP, DP and SF BVAP and %BVAP\*

SLDL	Total VAP DP	Total VAP SF	BNH VAP DP	BNH VAP SF	DP % BNH	SF % BNH
1	36,191	35,951	5,234	4,990	14.5%	13.9%
2	35,429	35,725	1,238	1,375	3.5%	3.8%
3	35,845	35,849	8,205	8,171	22.9%	22.8%
4	34,252	34,151	4,194	4,138	12.2%	12.1%
5	34,407	34,435	4,355	4,209	12.7%	12.2%
6	35,108	35,051	6,295	6,660	17.9%	19.0%
7	34,438	34,547	1,350	1,349	3.9%	3.9%
8	34,289	34,257	6,164	6,331	18.0%	18.5%
9	34,194	34,332	629	632	1.8%	1.8%
10	34,739	34,750	5,683	5,491	16.4%	15.8%
11	34,327	34,144	81	148	0.2%	0.4%
12	34,722	34,821	562	509	1.6%	1.5%
13	34,843	34,775	1,906	2,040	5.5%	5.9%
14	35,007	35,083	847	841	2.4%	2.4%
15	33,496	33,420	3,882	3,920	11.6%	11.7%
16	35,164	35,178	4,103	4,108	11.7%	11.7%
17	35,152	35,160	1,358	1,435	3.9%	4.1%
18	34,843	34,775	1,882	1,919	5.4%	5.5%
19	<b>34,527</b>	<b>34,604</b>	<b>19,647</b>	<b>19,787</b>	<b>56.9%</b>	<b>57.2%</b>
20	35,430	35,371	1,406	1,265	4.0%	3.6%
21	35,255	35,261	3,084	3,090	8.7%	8.8%
22	34,758	34,827	2,042	1,923	5.9%	5.5%
23	35,595	35,625	1,244	1,308	3.5%	3.7%
24	34,457	34,430	416	485	1.2%	1.4%
25	32,222	32,170	4,791	5,097	14.9%	15.8%
26	33,112	33,195	355	459	1.1%	1.4%
27	34,997	35,226	466	518	1.3%	1.5%
28	35,631	35,588	9,827	9,929	27.6%	27.9%
29	34,716	34,708	976	1,051	2.8%	3.0%
30	34,602	34,651	1,495	1,386	4.3%	4.0%
31	35,481	35,541	5,862	5,982	16.5%	16.8%
32	<b>35,514</b>	<b>35,496</b>	<b>17,842</b>	<b>17,750</b>	<b>50.2%</b>	<b>50.0%</b>
33	35,029	34,950	8,727	8,756	24.9%	25.1%
34	34,268	34,231	537	537	1.6%	1.6%
35	34,864	34,868	5,105	5,048	14.6%	14.5%
36	35,495	35,584	4,425	4,621	12.5%	13.0%
37	35,564	35,483	9,152	9,189	25.7%	25.9%
38	34,595	34,584	6,869	6,953	19.9%	20.1%
39	35,561	35,635	1,811	1,791	5.1%	5.0%
40	35,695	35,507	4,523	4,673	12.7%	13.2%
41	33,850	33,802	3,975	3,894	11.7%	11.5%
42	34,184	34,156	4,281	4,100	12.5%	12.0%
43	34,074	34,167	2,234	2,196	6.6%	6.4%
44	33,870	33,824	3,873	3,823	11.4%	11.3%
45	34,324	34,370	5,059	5,225	14.7%	15.2%
46	34,280	34,103	2,531	2,520	7.4%	7.4%
47	34,913	34,968	5,970	6,161	17.1%	17.6%
48	34,954	34,971	1,988	2,008	5.7%	5.7%
49	33,460	33,728	4,088	4,301	12.2%	12.8%
50	34,982	34,949	3,084	2,983	8.8%	8.5%
51	34,795	34,669	1,831	1,872	5.3%	5.4%
52	<b>36,021</b>	<b>35,997</b>	<b>21,398</b>	<b>21,262</b>	<b>59.4%</b>	<b>59.1%</b>
53	35,882	35,861	18,146	17,966	50.6%	50.1%

Source: 20201116 PPMF, \* Bolded Districts are Black Majority-Minority Districts

Table 3c4b: State Legislative Districts DP and SF VAP, DP and SF BVAP and %BVAP\* Continued

SED	Total VAP DP	Total VAP SF	BNH VAP DP	BNH VAP SF	DP % BNH	SF % BNH
54	35,877	35,922	20,205	20,393	56.3%	56.8%
55	35,716	35,844	26,271	26,125	73.6%	72.9%
56	33,999	34,006	21,101	20,878	62.1%	61.4%
57	34,437	34,552	19,691	19,770	57.2%	57.2%
58	32,951	32,947	19,782	19,752	60.0%	60.0%
59	32,817	32,765	24,282	24,221	74.0%	73.9%
60	35,563	35,616	23,179	23,316	65.2%	65.5%
61	34,810	34,846	8,244	8,139	23.7%	23.4%
62	34,028	34,089	6,044	5,813	17.8%	17.1%
63	39,623	39,599	4,797	5,031	12.1%	12.7%
64	34,413	34,366	5,089	5,103	14.8%	14.8%
65	34,489	34,355	11,713	11,613	34.0%	33.8%
66	35,403	35,447	8,755	8,669	24.7%	24.5%
67	33,198	33,259	21,853	21,668	65.8%	65.1%
68	34,003	34,105	17,121	17,278	50.4%	50.7%
69	34,638	34,576	21,464	21,188	62.0%	61.3%
70	35,683	35,788	18,624	18,702	52.2%	52.3%
71	34,599	34,436	19,960	19,949	57.7%	57.9%
72	34,190	34,176	21,258	21,086	62.2%	61.7%
73	33,495	33,133	3,492	3,214	10.4%	9.7%
74	35,918	35,887	8,371	8,423	23.3%	23.5%
75	34,918	34,878	8,810	8,739	25.2%	25.1%
76	33,358	33,396	25,385	25,639	76.1%	76.8%
77	35,222	35,198	20,573	20,486	58.4%	58.2%
78	33,663	33,695	21,812	21,674	64.8%	64.3%
79	37,893	37,981	5,675	5,262	15.0%	13.9%
80	34,186	34,048	5,970	5,992	17.5%	17.6%
81	35,995	35,958	7,272	7,389	20.2%	20.5%
82	36,248	36,488	19,640	19,962	54.2%	54.7%
83	33,707	33,681	17,383	17,539	51.6%	52.1%
84	35,643	35,558	18,334	18,070	51.4%	50.8%
85	34,617	34,625	14,511	14,560	41.9%	42.1%
86	34,898	34,967	5,782	5,746	16.6%	16.4%
87	35,215	35,155	2,939	2,980	8.3%	8.5%
88	33,404	33,512	5,477	5,899	16.4%	17.6%
89	35,254	35,283	10,821	10,739	30.7%	30.4%
90	35,342	35,286	12,112	11,908	34.3%	33.7%
91	34,034	34,066	5,089	5,071	15.0%	14.9%
92	34,962	34,899	4,040	4,048	11.6%	11.6%
93	34,805	34,765	5,561	5,386	16.0%	15.5%
94	35,074	35,076	2,452	2,612	7.0%	7.4%
95	37,081	37,023	1,557	1,532	4.2%	4.1%
96	34,250	34,301	3,969	3,768	11.6%	11.0%
97	33,792	33,762	18,563	18,683	54.9%	55.3%
98	33,998	33,888	18,921	18,863	55.7%	55.7%
99	33,832	33,812	20,345	20,211	60.1%	59.8%
100	33,234	33,064	4,702	4,716	14.1%	14.3%
101	35,634	35,854	5,904	5,876	16.6%	16.4%
102	33,010	33,320	3,074	3,158	9.3%	9.5%
103	32,747	32,667	18,773	18,898	57.3%	57.9%
104	34,167	34,257	4,470	4,588	13.1%	13.4%
105	33,688	33,596	4,110	3,778	12.2%	11.2%

Source: 20201116 PPMF, \* Bolded Districts are Black Majority-Minority Districts

Table 3c4c: State Legislative Districts # and % Difference in BVAP Between SF and DP\*

SLD	BNH VAP DP	BNH VAP SF	BNH VAP # Diff	BNH VAP % Diff
1	5,234	4,990	244	4.9%
2	1,238	1,375	-137	-10.0%
3	8,205	8,171	34	0.4%
4	4,194	4,138	56	1.4%
5	4,355	4,209	146	3.5%
6	6,295	6,660	-365	-5.5%
7	1,350	1,349	1	0.1%
8	6,164	6,331	-167	-2.6%
9	629	632	-3	-0.5%
10	5,683	5,491	192	3.5%
11	81	148	-67	-45.3%
12	562	509	53	10.4%
13	1,906	2,040	-134	-6.6%
14	847	841	6	0.7%
15	3,882	3,920	-38	-1.0%
16	4,103	4,108	-5	-0.1%
17	1,358	1,435	-77	-5.4%
18	1,882	1,919	-37	-1.9%
19	<b>19,647</b>	<b>19,787</b>	<b>-140</b>	<b>-0.7%</b>
20	1,406	1,265	141	11.1%
21	3,084	3,090	-6	-0.2%
22	2,042	1,923	119	6.2%
23	1,244	1,308	-64	-4.9%
24	416	485	-69	-14.2%
25	4,791	5,097	-306	-6.0%
26	355	459	-104	-22.7%
27	466	518	-52	-10.0%
28	9,827	9,929	-102	-1.0%
29	976	1,051	-75	-7.1%
30	1,495	1,386	109	7.9%
31	5,862	5,982	-120	-2.0%
32	<b>17,842</b>	<b>17,750</b>	<b>92</b>	<b>0.5%</b>
33	8,727	8,756	-29	-0.3%
34	537	537	0	0.0%
35	5,105	5,048	57	1.1%
36	4,425	4,621	-196	-4.2%
37	9,152	9,189	-37	-0.4%
38	6,869	6,953	-84	-1.2%
39	1,811	1,791	20	1.1%
40	4,523	4,673	-150	-3.2%
41	3,975	3,894	81	2.1%
42	4,281	4,100	181	4.4%
43	2,234	2,196	38	1.7%
44	3,873	3,823	50	1.3%
45	5,059	5,225	-166	-3.2%
46	2,531	2,520	11	0.4%
47	5,970	6,161	-191	-3.1%
48	1,988	2,008	-20	-1.0%
49	4,088	4,301	-213	-5.0%
50	3,084	2,983	101	3.4%
51	1,831	1,872	-41	-2.2%
52	<b>21,398</b>	<b>21,262</b>	<b>136</b>	<b>0.6%</b>
53	<b>18,146</b>	<b>17,966</b>	<b>180</b>	<b>1.0%</b>

Source: 20201116 PPMF, \* Bolded Districts are Black Majority-Minority Districts



Table 3c4c: State Legislative Districts # and % Difference in BVAP Between SF and DP\* Continued

SLDL	BNH VAP DP	BNH VAP SF	BNH VAP # Diff	BNH VAP % Diff
54	20,205	20,393	-188	-0.9%
55	26,271	26,125	146	0.6%
56	21,101	20,878	223	1.1%
57	19,691	19,770	-79	-0.4%
58	19,782	19,752	30	0.2%
59	24,282	24,221	61	0.3%
60	23,179	23,316	-137	-0.6%
61	8,244	8,139	105	1.3%
62	6,044	5,813	231	4.0%
63	4,797	5,031	-234	-4.7%
64	5,089	5,103	-14	-0.3%
65	11,713	11,613	100	0.9%
66	8,755	8,669	86	1.0%
67	21,853	21,668	185	0.9%
68	17,121	17,278	-157	-0.9%
69	21,464	21,188	276	1.3%
70	18,624	18,702	-78	-0.4%
71	19,960	19,949	11	0.1%
72	21,258	21,086	172	0.8%
73	3,492	3,214	278	8.6%
74	8,371	8,423	-52	-0.6%
75	8,810	8,739	71	0.8%
76	25,385	25,639	-254	-1.0%
77	20,573	20,486	87	0.4%
78	21,812	21,674	138	0.6%
79	5,675	5,262	413	7.8%
80	5,970	5,992	-22	-0.4%
81	7,272	7,389	-117	-1.6%
82	19,640	19,962	-322	-1.6%
83	17,383	17,539	-156	-0.9%
84	18,334	18,070	264	1.5%
85	14,511	14,560	-49	-0.3%
86	5,782	5,746	36	0.6%
87	2,939	2,980	-41	-1.4%
88	5,477	5,899	-422	-7.2%
89	10,821	10,739	82	0.8%
90	12,112	11,908	204	1.7%
91	5,089	5,071	18	0.4%
92	4,040	4,048	-8	-0.2%
93	5,561	5,386	175	3.2%
94	2,452	2,612	-160	-6.1%
95	1,557	1,532	25	1.6%
96	3,969	3,768	201	5.3%
97	18,563	18,683	-120	-0.6%
98	18,921	18,863	58	0.3%
99	20,345	20,211	134	0.7%
100	4,702	4,716	-14	-0.3%
101	5,904	5,876	28	0.5%
102	3,074	3,158	-84	-2.7%
103	18,773	18,898	-125	-0.7%
104	4,470	4,588	-118	-2.6%
105	4,110	3,778	332	8.8%
Grand Total	902,350	902,278	72	0.0%

Source: 20201116 PPMF, \* Bolded Districts are Black Majority-Minority Districts

Please note: these figures above differ from those in State Legislative Issues (below). The data in the figures above are for the legislative boundaries as they existed after the 2010 based redistricting. The data in State Legislative Issues (below) are for the legislative boundaries as they existed prior to the 2010 Census. We use this latter definition because these are the data exactly as were published by IPUMS.

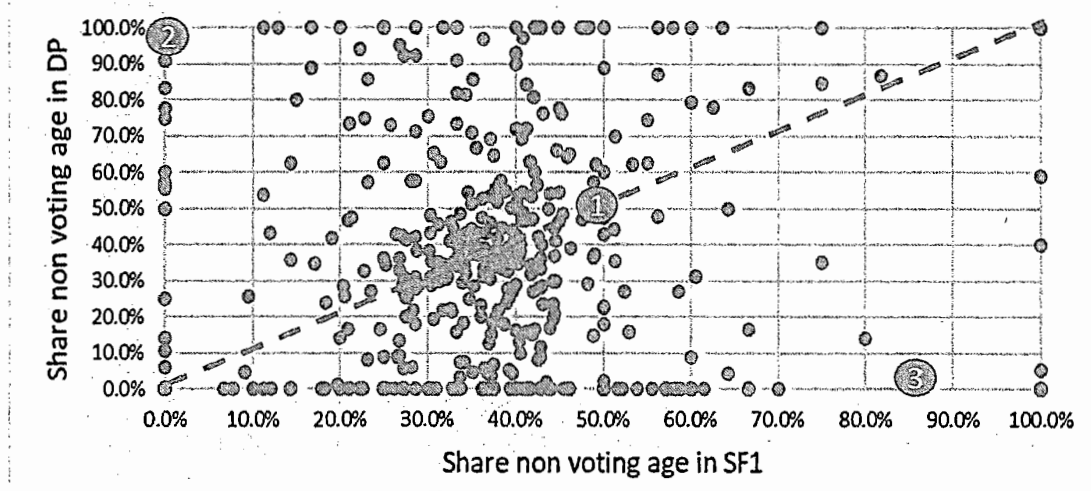
### **3c5: Summary Statistics and Analysis at Different levels of Geography**

Each layer of geography we examine has a unique importance. With high levels of geography such as Congressional Districts, we find numeric and percent differences that are small by demographic standards but hold *significant* importance in redistricting where precision to the individual is required by law. At other levels of geography, such as places, unified school districts, and legislative districts, there are geographies that have very small numeric errors, but large percent errors (typically in small places). And there are other geographies that have very small percent errors, but large numeric errors (typically in large places). In order to quickly identify geographies that had the most significant errors, we deployed a statistical technique called "loss functions." In mathematical optimization and decision theory, a loss function or cost function is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event. (Hough and Swanson, 2006).

The analytic tables illustrate geographies with the largest differences for VAP and NVA. These tables are followed in turn with scatterplot figures showing the change in percent share of NVA from the SF to DP files. Non-Voting Age (NVA) children represent some share of the total population at every level of geography—sometimes more, sometimes less, but usually around 20% total. In theory, if children make up a share of the total in the SF data, it should be approximately the same in the DP data.

In the scatterplot below, the SF data are on the X axis, and the DP data are on the Y axis. The further the data are from the orange, dotted line, the greater the departure from the SF data. Exceptional cases of "0%" literally reflect NVA populations going from none to some, or some to none.

Figure 3c2: Illustrative Example of Scatterplot of % NVA in SF and % NVA in DP



- ① The orange dot representing data at position 1 are what we would expect. This shows that 50% of the population are children in the SF file (original census data shown on the X axis) and that 50% of the population are also 50% children in the DP file (shown on the Y axis). The further the blue data points are from the orange line, the greater the difference between the SF and DP % children values.
- ② The orange dot representing data at position 2 is *not* what we would expect. This shows that 0% of the population are children in the SF file (original census data shown on the X axis) but are 100% of the population in the DP file (shown on the Y axis). This is practically implausible.
- ③ The orange dot representing data at position 3 is also *not* what we would expect. This shows that 100% of the population are children in the SF file (original census data shown on the X axis) but are 0% of the population in the DP file (shown on the Y axis). This is also practically implausible.

An important component of analyzing and interpreting the impact of DP is knowing the number of pieces of geography there are at each level (see Figure 5d4 below). Experience with the existing data tell us that geographic layers with more pieces and smaller pieces, such as Places, harbor more data issues than layers with fewer, larger pieces, such as Congressional Districts. In the following analysis, we examine the numeric and percent difference in VAP, then NVA side-by-side by race for counties, state legislative districts, Unified School Districts; and Alabama Places. This is followed by a series of detailed cases using census blocks

**Figure 3c5a: Number of Geographic Units in the US and Alabama**

	<u>U.S.</u>	<u>Alabama</u>
Counties	3,142	67
Congressional Districts	435	7
State Legislative Districts		105
State Senate Districts		35
Unified School Districts	10,914	134
Places	29,514	578
Census Blocks	11,155,486	252,266

Based on data from 2010 Census Summary File

**County Issues**

We proceed here with an analysis of differences at the county level. There is only one county with notable differences among Black / African American VAP and NVA – DeKalb. The Asian population has many more significant differences. In Tallapoosa County, all 63 NVA children from the SF file are removed, leaving no Asian children there. This happens to 17 other counties in Alabama as well and can be seen in Table 5d3. This difference can be compared with almost equal and dramatic increases of Asian children in other counties such as Franklin County. How would each county deal with the complete absence or introduction of a minority population erroneously? Similarly, in Monroe and Pickens Counties, how would they manage the reporting of numerous Hispanic children that don’t actually exist? Would they be compelled to provide language programs and services, and be at risk of violating the law for not providing for a population that didn’t actually exist?

<b>Black / African American</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
DeKalb	-122, +83	-15%, +29%

<b>Asian</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
Autauga County	+151,-126	+44%, -98%
Franklin County	-36, +59	-80%, +328%
Lauderdale County	-100, +82	-17%, +75%
St. Clair County	+44, -108	+12%, -78%
Tallapoosa County	+53, -63	+40%, -100%

<b>Hispanic</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
Monroe	-74, +79	-53%, +99%
Pickens	-87, +74	-36%, +104%

In the scatterplots by race below, the SF data are on the X axis, and the DP data are on the Y axis.

Figure 3c5a: Asian Differences in NVA between SF and DP by Alabama County

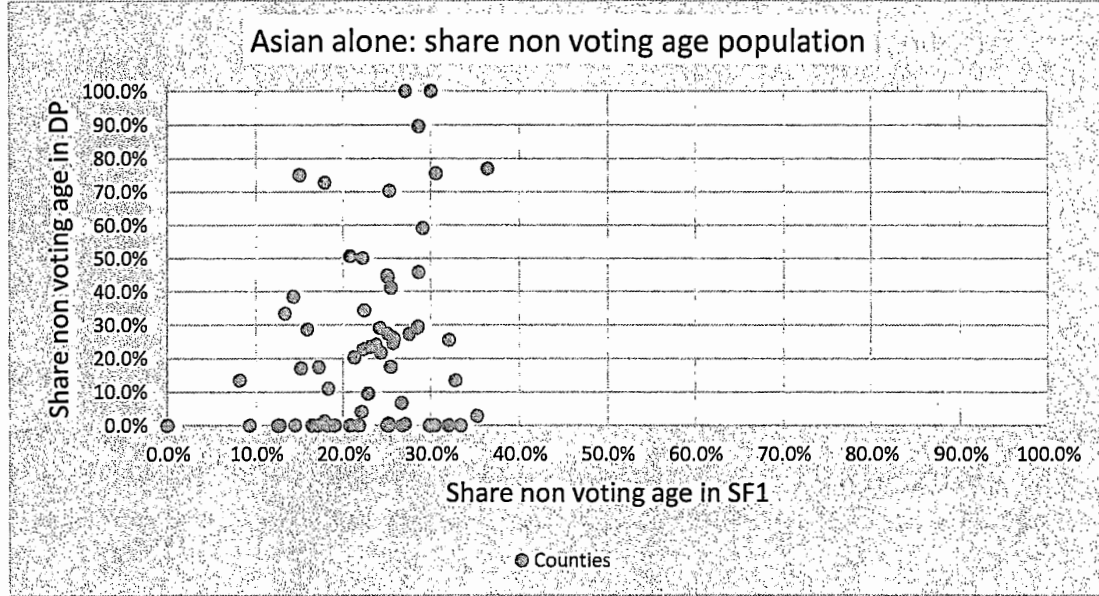
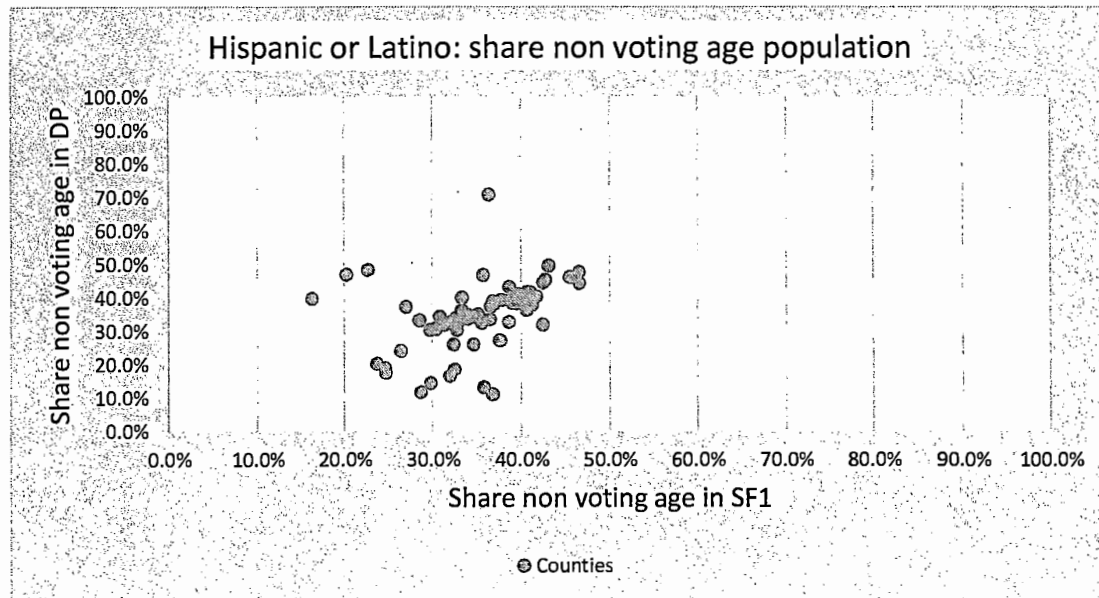


Figure 3c5b: Hispanic Differences in NVA between SF and DP by Alabama County



**State Legislative Issues**

We continue by assessing 105 state legislative districts. Please note: these figures differ from those in Figure 5d3. The data here are for the legislative boundaries as they existed prior to the 2010 Census. The data in Figure 5d3 are for the legislative boundaries as they existed after the 2010 Census redistricting. The districts in this analysis are what we refer to as being "off-spine." That is, the DP process does not make a direct explicit effort to control or manage data at this level of geography. Its results are a function of smaller geographies such as blocks that comprise it.

Differences in VAP and NVA by Legislative District and Race in Alabama (SF – PL data).

For Black / African Americans, there are six districts with both significant numeric and percent differences, which would result in a *significant* change in demographic complexion in these areas.

<b>Black / African Americans</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
District 25	-376, +199	-5%, +7%
District 35	+414, -73	+8%, -4%
District 62	+421, -451	+5%, -12%
District 64	-262, +440	-3%, +18%
District 68	-93, -533	-1%, -8%
District 70	-361, +287	-2%, +4%

For the Asian population, which is smaller than the Black / African American population, the numeric changes are small, but the percent changes are large. As with counties, there are whole districts where the Asian NVA children population are either wiped out (such as District 36, see figure 5d7) or appear out of nowhere. In some districts, the differences are compounding (both NVA and VAP gain or lose population); in others they are offsetting.

<b>Asian</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
District 19	-27, +133	-11%, +359%
District 32	+48, +41	+56%, +456%
District 36	+67, -96	+21%, -100%
District 61	-112, +75	-54%, +117%
District 88	+141, -125	+33%, -87%
District 102	-31, +87	-27%, +242%

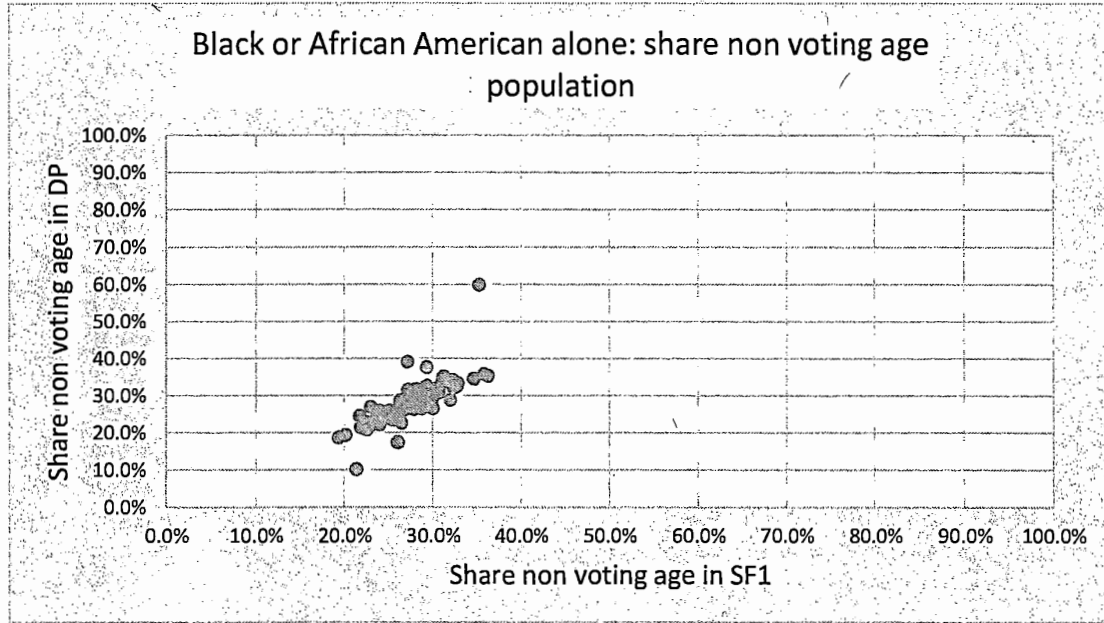
In cases such as District 36, the interpretation is that DP eliminated all 96 of the Asian NVA children, hence -100%.

For the Hispanic population, similar to the Black / African American population, there are districts with large, severe changes that fundamentally would change the demographic complexion of these districts.

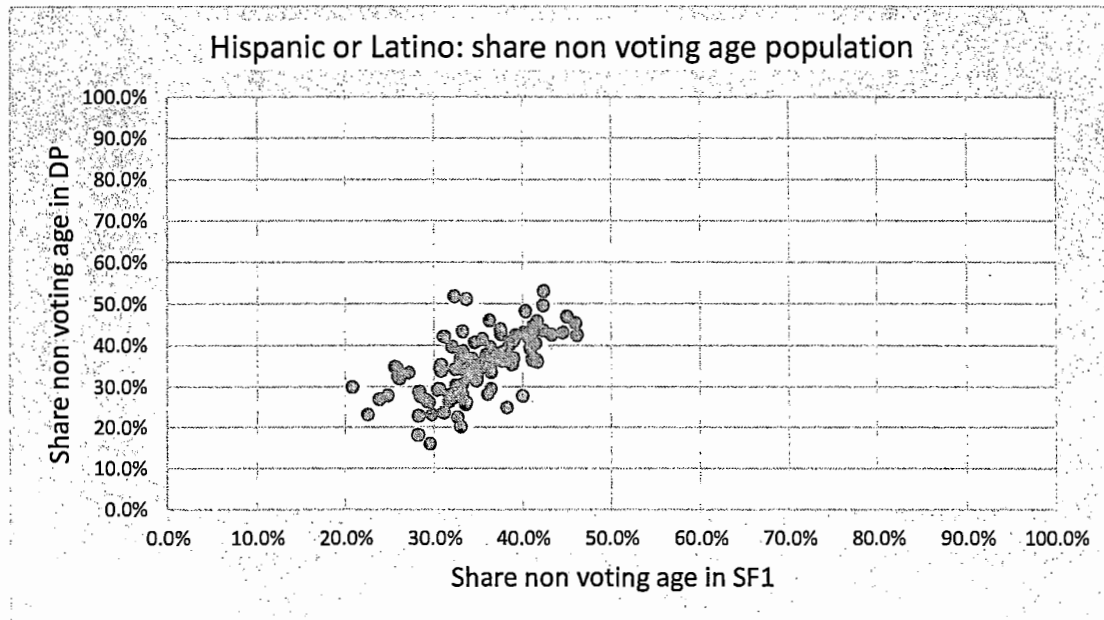
<b>Hispanic</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
District 57	-173, +160	-28%, +50%
District 83	+199, -205	-19%, -39%
District 87	+180, -192	+19%, -31%
District 94	+64, -239	+4%, -28%

In the scatterplots by race below, the SF data are on the X axis, and the DP data are on the Y axis.

**Figure 3c5c: Black Differences in NVA between SF and DP by Alabama Legislative District**

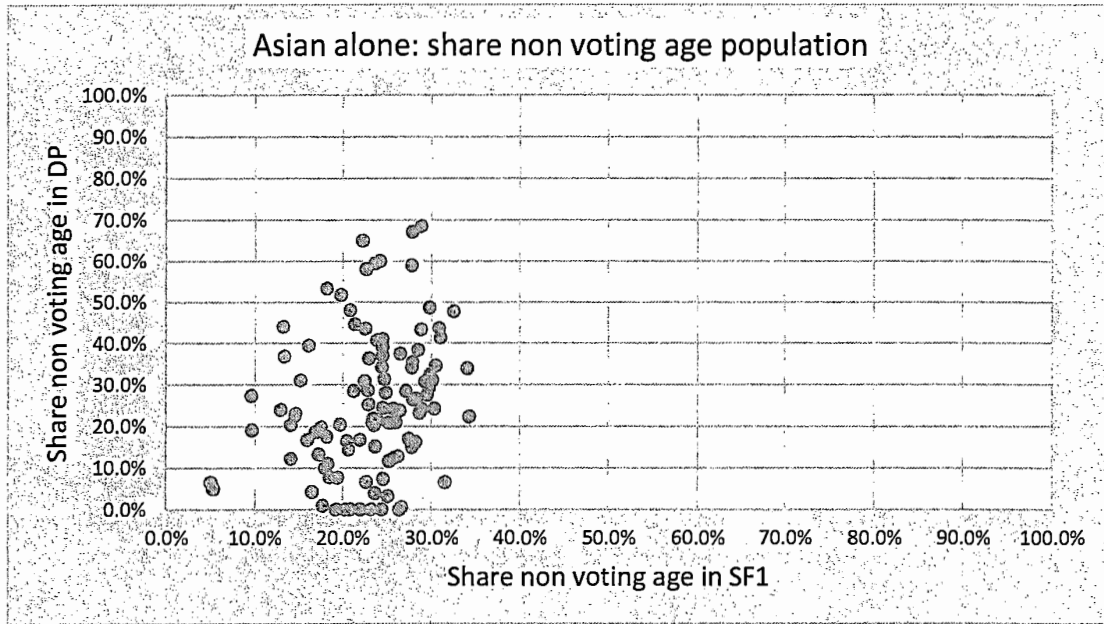


**Figure 3c5d: Hispanic Differences in NVA between SF and DP by Alabama Legislative District**



In the scatterplots by race below, the SF data are on the X axis, and the DP data are on the Y axis.

**Figure 3c5e: Asian Differences in NVA between SF and DP by Alabama Legislative District**





**Unified School District Issues**

There are 134 unified school districts in Alabama – slightly more than the number of legislative districts. There are also “off-spine” geography, subject to more variation in the analysis because no effort has been made to “balance” the size of these school district populations. They vary from small to very large.

As with legislative districts, the Black / African American population varies significantly with both large numeric and percent differences in districts such as Hoover, Mountain Brook City, and Talladega County.

<b>Black</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
Hoover City	-225, +312	-3%, +8%
Auburn City	+342, -179	+5%, -8%
Mountain Brook City CSD	+70, +148	+56%, +185%
Pike County SD	+123, -238	+3%, -18%
Talladega County SD	-317, +173	-3%, +5%

As seen in other levels of geography, the relatively small Asian population is dramatically affected. In a pattern seen with other minority groups, the Asian NVA population is frequently “zeroed out”.

<b>Asian</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
Autauga County SD	+151, -126	+44%, -98%
Bessemer County SD	+103, +16	+240%, +160%
Fort Payne SD	-63, +61	-83%, +156%
Jefferson CSD	-205, +172	-17%, +54%

For the Hispanic population, the errors are most numerous. It is difficult to imagine how a school district would manage providing (or not providing) services and support to hundreds of minority students that were reported to exist and didn’t, or vice versa.

<b>Hispanic</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
Athens SD	-137, +196	-12%, +25%
Eufaula CSD	-115, +110	-31%, +56%
Limestone SD	+140, -203	+9%, -19%
Madison SD	-65, +274	-3%, +21%
Midfield City CSD	-48, +61	-87%, +277%
Monroe CSD	-74, +79	-53%, +99%
Pickens CSD	-87, +74	-36%, +104%
Vestavia CSD	+177, -154	+34%, -49%

In the scatterplots by race below, the SF data are on the X axis, and the DP data are on the Y axis.

Figure 3c5f: Black Differences in NVA between SF and DP by Alabama USD

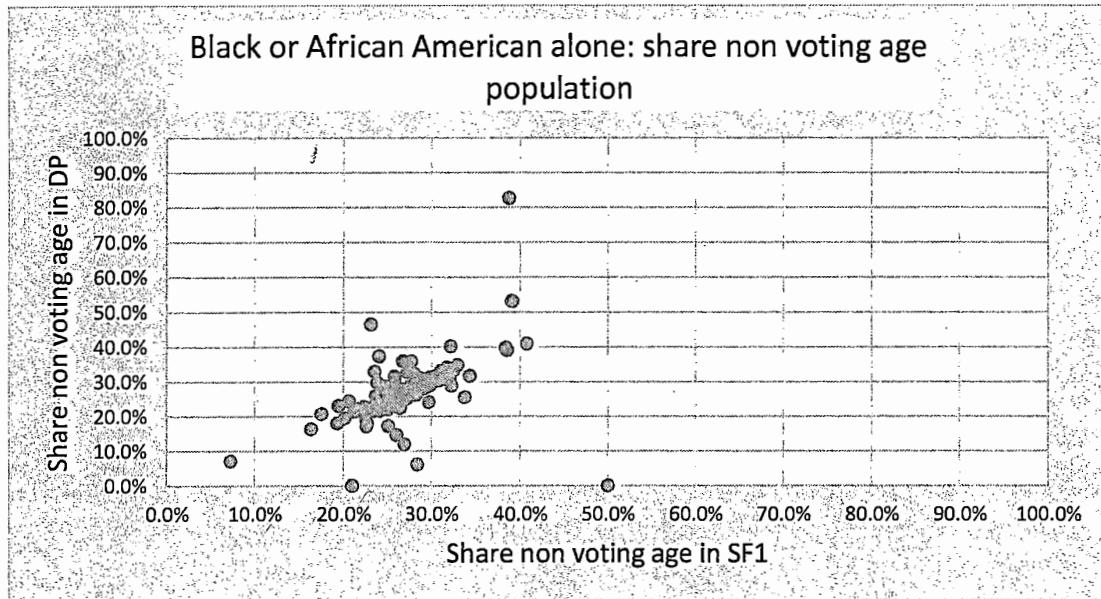
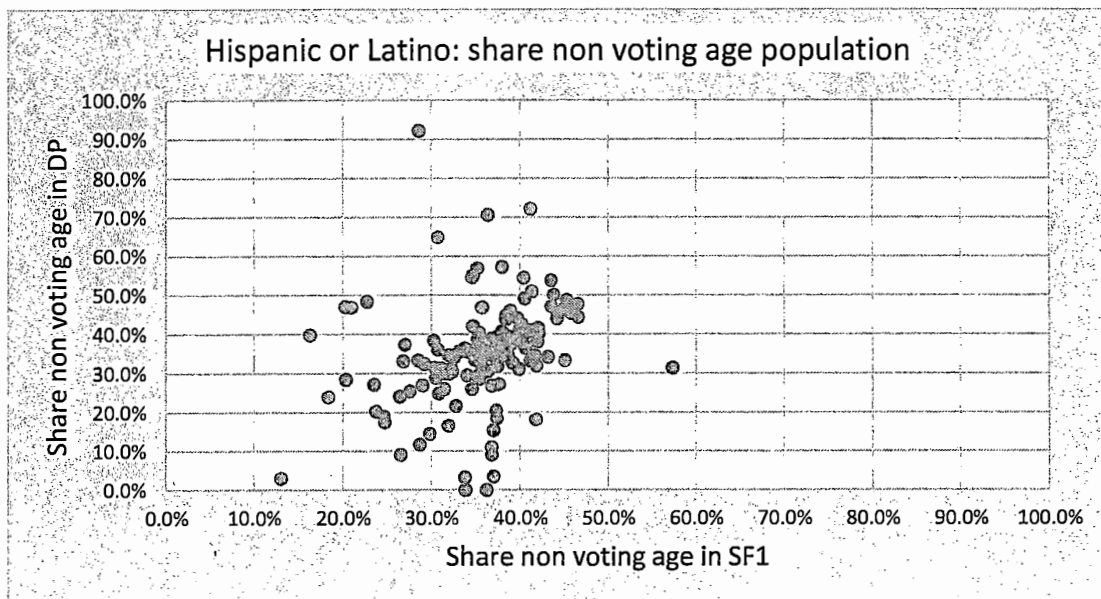
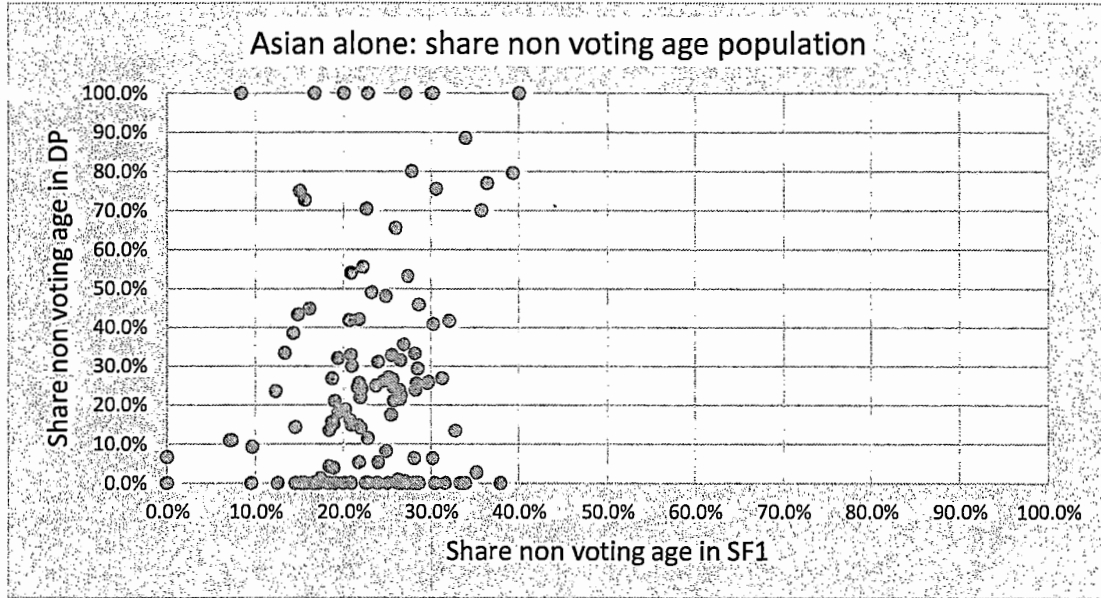


Figure 3c5g: Hispanic Differences in NVA between SF and DP by Alabama USD



In the scatterplots by race below, the SF data are on the X axis, and the DP data are on the Y axis.

Figure 3c5h: Asian Differences in NVA between SF and DP by Alabama USD



**Place Issues**

There are 578 incorporated and Census Designated Places in Alabama. Many of them are small, and differential privacy often has a big impact on the accuracy of data for small places. In the 2010 Census, 272 of the 578 had a population of less than 1,000 people, and 514 of the 578 had a population of less than 10,000 people.

Some of the changes based on application of differential privacy were large. For example, the 2010 Census reported a total population of 215 in Belk Town, but after differential privacy was applied, that number was changed to 153, which amounts to a 29 percent change. In Graysville City, the 2010 Census reported a population of 2,165, but after differential privacy was applied, the number was changed to 2,043, a loss of 122 people.

For large cities, an error of 122 people might not make much difference. But in smaller places, like most of those in Alabama, such a distortion can have a big impact. With numerous places, and the largest differentials in size of all the geographies we examine, places show the greatest errors.

Not surprisingly, the Black / African American population shows numerous significant errors. While some of the bases are small, the percent differences for places like Carlton and Peterman are impressive. looioioo0ooiii976i7Black / African American NVA are “zeroed out” of an amazing 68 places under DP.

Differences in VAP and NVA by Place and Race in Alabama (SF – PL data)

<b>Black</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
Carlton CDP	-19, +48	-90%, +4,800%
Foley city	-279, +137	-18%, +21%
Mountain Brook city	+99, +175	+77%, +213%
Peterman CDP	+61, +16	+1,017%, +800%
Prattville city	-368, +236	-10%, +13%
Weaver city	-141, +110	-53%, +76%

As with other layers of geography, Asians are dramatically affected by DP at the place level. Aside from significant numeric and percent errors, their NVA population is “zeroed out” of an amazing 131 places.

<b>Asian</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
Bessemer city	+103, +16	+225%, +240%
Pelham city	-139, +133	-37%, +95%
Prattville city	+104, -105	+29%, -86%

Finally, Hispanics are also dramatically affected by DP at the place level. Aside from significant numeric and percent errors, the Hispanic NVA population is “zeroed out” of an alarming 120 places.

<b>Hispanic</b>	<b># Error VAP, NVA</b>	<b>% Error, VAP, NVA</b>
Athens city	-137, +196	-12%, +25%
Coker town	+79, -7	+878%, -78%
Eufaula city	-115, +110	-31%, +56%
Midfield city	-48, +61	-87%, +277%
Vestavia Hills city	+177, -154	+34%, -49%

In the scatterplots by race below, the SF data are on the X axis, and the DP data are on the Y axis.

Figure 3c5i: Black Differences in NVA between SF and DP by Alabama Place

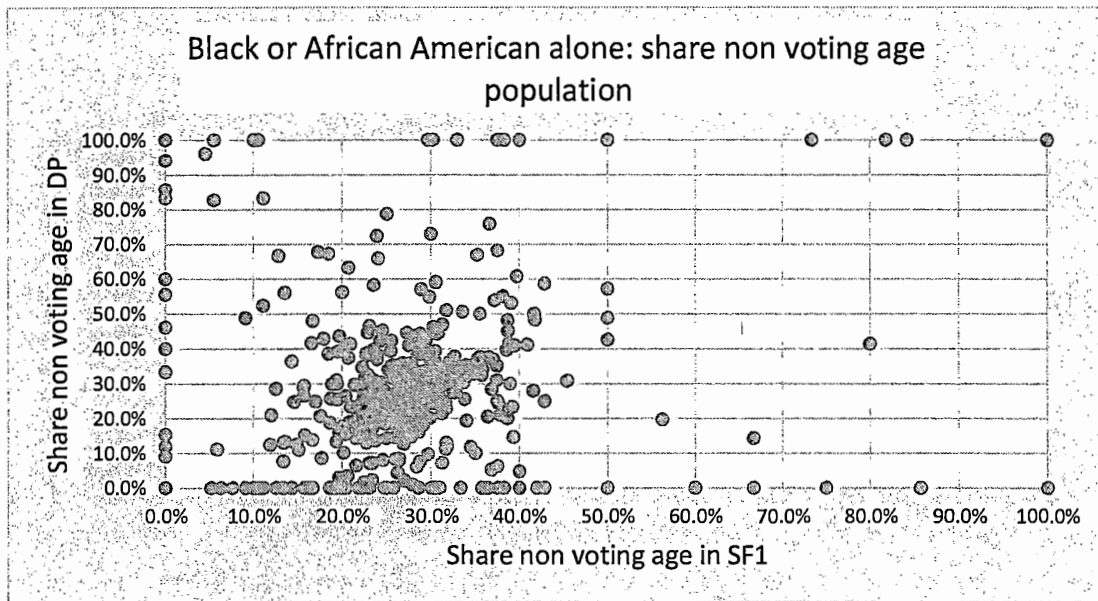
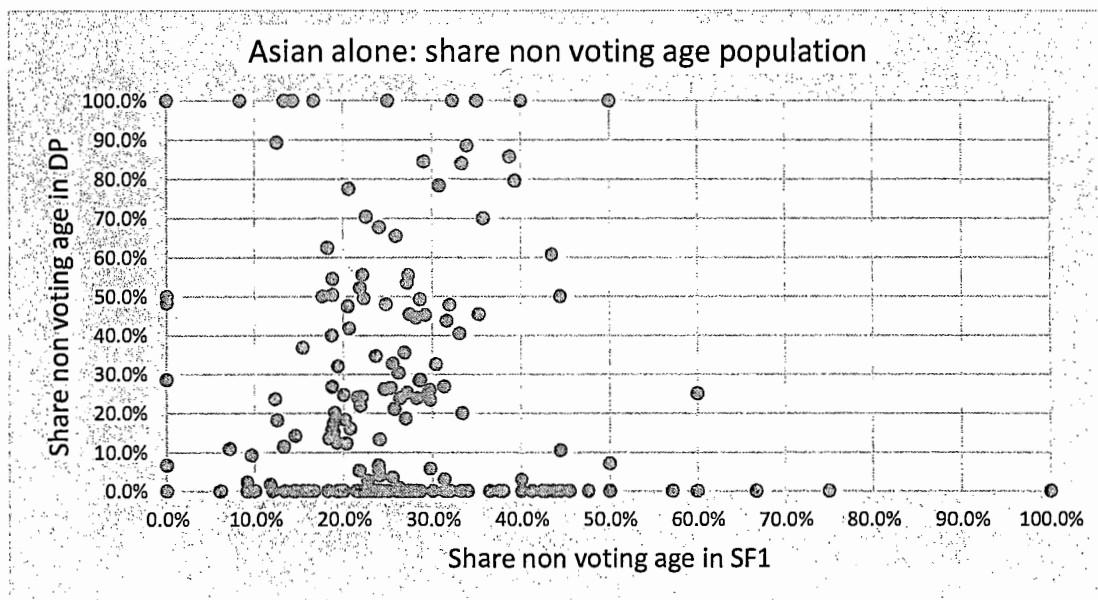
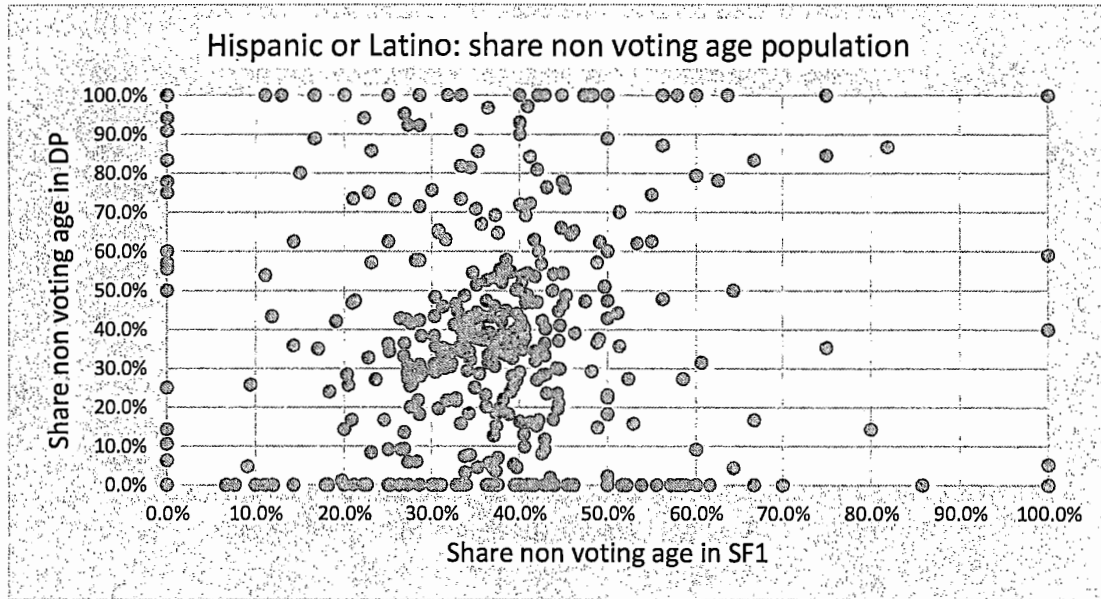


Figure 3c5j: Asian Differences in NVA between SF and DP by Alabama Place



In the scatterplots by race below, the SF data are on the X axis, and the DP data are on the Y axis.

**Figure 3c5k: Hispanic Differences in NVA between SF and DP by Alabama Place**



#### **Section 4: Summary and Conclusions**

The evidence we have from the DP demonstration products is that the output exhibits a degree of statistical adjustment that irreparably harms Census data in different ways at all levels of spine and off-spine geography. Ruggles et al. (2019: 403-404) observe that differential privacy is inconsistent with the statutory obligations, history, and core mission of the Census Bureau and that by imposing unrealistic rules, the Census Bureau may be forced to “lock up” data that are indispensable for basic research and policy research. To this, we add that the errors introduced by DP would adversely affect political redistricting and the demographics industry, as well as planning for public health and public safety needs, and the planning needs of educational, municipal, and regional planning organizations.

As we noted early in this report, Ruggles et al. (2019: 406) also observe that, because it focuses on concealing individual characteristics instead of respondent identities, DP is a blunt and inefficient instrument for disclosure control. In regard to how blunt DP can be, recall the examples of median and extreme income given in Section 3 (p. 18) and consider the fact that the Census Bureau will not be able to alter DP levels for each individual query. Instead, it will set universal levels within given domains. In terms of the income example, this implies that a query about median income will run into the same level of inaccuracy as a query on maximum income, even though the probability that the query about median income could result in “leaked” information is so small as to be virtually zero.

While the threat of a confidentiality breach is always present, the Census Bureau has not reported any such breaches from prior census data releases, a fact also noted by Ruggles et al. (2019: 404) who state, “... [T]here is not a single documented case of anyone outside The Census Bureau revealing the responses of a particular identified person in public use Decennial Census or ACS data.” These facts suggest that the Census Bureau’s current confidentiality and privacy protocols are effective. We understand that new threats can emerge, but as was the case with the “handheld” devices that led to the 2010 FDCA debacle, we believe that DP is still an immature technology in the hands of an agency with insufficient experience to implement it in a manner that will preserve the accuracy of small area data while protecting the privacy and confidentiality of respondent information. Far more testing and development needs to be done, allowing both Bureau staff and stakeholders to become familiar with DP in order to make reasonable decisions about using it to statistically adjust the 2020 Census.

While the Census Bureau has invested in the deployment of the 2020 Census using DP, we conclude that DP should not be applied to the 2020 Decennial Census and related products. The results of our analysis and the observations of others strongly suggest that the deployment of DP will violate the mission of the Census Bureau of publishing quality, accurate data available to the public, to the extent that the data as published under DP would not allow states to comply with the law (starting with redistricting). Perfect compliance with Title XIII and perfect privacy would mean that no census data be released at all. In the absence of such standards, the Bureau is inventing implausible and oftentimes demographically impossible data at the 11<sup>th</sup> hour that are demonstrably flawed and would provide a disservice to a wide variety of census data consumers. In the absence of such standards, and until DP is adequately vetted and standards exist for balancing privacy and quality, we conclude that the US Census Bureau should continue using their established DAS methodology from the 2010 Census. The Census Bureau has this methodology “on the shelf” and should have immediate access to sufficient human capital in the form of staff and contractor experience required to use it in a short period of time.

### **Appendix 1: Differential Privacy Data**

In June 2020, the Census Bureau announced plans to release a Privacy-Protected Microdata File (**PPMF**) after each programming sprint, for which the Bureau generates a corresponding set of quality metrics. The Bureau is continually modifying its differentially private algorithm, and each version of the PPMF will reflect those modifications. Data users may use the PPMF to track changes in accuracy and utility. To make these data more user-friendly, IPUMS NHGIS is creating a Privacy-Protected Summary File (**PPSF**) from each version of the PPMF. Our PPMF consists of tabulations where each row represents a geographic unit and each column represents a summary statistic (e.g., VAP population in DP and SF).

To facilitate comparisons, we (NHGIS) link comparable data from the PPSF and original 2010 Census Summary File 1. These linked files comprise the IPUMS NHGIS Privacy-Protected 2010 Census Demonstration Data product<sup>13</sup>.

### **Appendix 2: Terms**

(DP) Differential Privacy: A statement by Cynthia Dwork (2006): "A statistic is a quantity computed from a sample. If a database is a representative sample of an underlying population, the goal of a privacy-preserving statistical database is to enable the user to learn properties of the population as a whole, while protecting the privacy of the individuals in the sample." Since the application of differential privacy occurs within the Census Bureau's Disclosure Avoidance Systems (DAS), that term has sometimes been used to describe the use of differential privacy. To avoid confusion, the term differential privacy (DP) is used here to distinguish the version of DAS that includes DP from other versions of DAS.

(DAS) Disclosure Avoidance System: Before the Census Bureau publishes any statistic, they apply safeguards that help prevent someone from being able to trace that statistic back to a specific respondent. They call these safeguards "disclosure avoidance," although these methods are also known as "statistical disclosure controls" or "statistical disclosure limitations." Although it might appear that a published table shows information about a specific individual, the Census Bureau has taken steps to disguise the original data in such a way that the results are still useful. These steps include using statistical methods such as "data swapping" and "noise injection."<sup>14</sup>

(PPDD) Privacy-Protected Demonstration Data: To protect the confidentiality of 2020 Census respondents, the U.S. Census Bureau plans to use a framework termed "differential privacy." Beginning in October 2019, the Census Bureau began releasing privacy-protected demonstration data products (PPDD) to help users assess the impact of differential privacy on the utility and accuracy of Decennial Census data. This product was a differentially private version of the 2010 Decennial Census.

(PPMF) Privacy-Protected Microdata File: PPMFs are the underlying microdata files for the entire nation used to generate Detailed Summary Metrics.

(PPSF) Privacy-Protected Summary File: Produced by IPUMS NHGIS from each version of the PPMF.

(SF) Summary Files: 2010 census data as they were originally published by the Census Bureau.

(GQ) Group Quarters

(VAP) Voting Age Population >18

(NVA) Non-Voting Age < 18

---

<sup>13</sup> <https://www.nhgis.org/privacy-protected-demonstration-data#purpose>

<sup>14</sup> [https://www.census.gov/about/policies/privacy/statistical\\_safeguards.html](https://www.census.gov/about/policies/privacy/statistical_safeguards.html)



**Appendix 3: 2010 – 2019 Total Population Estimated Changes in Alabama**

Label	Total 2010	Total 2019	# Difference	% Difference
Total:	4,785,298	4,903,185	117,887	2%
Male:	2,322,243	2,369,611	47,368	2%
Under 18 years:	582,172	556,757	-25,415	-4%
Native	572,304	549,178	-23,126	-4%
Foreign born:	9,868	7,579	-2,289	-23%
Naturalized U.S. citizen	2,245	1,620	-625	-28%
Not a U.S. citizen	7,623	5,959	-1,664	-22%
18 years and over:	1,740,071	1,812,854	72,783	4%
Native	1,662,134	1,734,786	72,652	4%
Foreign born:	77,937	78,068	131	0%
Naturalized U.S. citizen	18,846	33,600	14,754	78%
Not a U.S. citizen	59,091	44,468	-14,623	-25%
Female:	2,463,055	2,533,574	70,519	3%
Under 18 years:	553,256	528,840	-24,416	-4%
Native	544,673	520,180	-24,493	-4%
Foreign born:	8,583	8,660	77	1%
Naturalized U.S. citizen	2,056	2,882	826	40%
Not a U.S. citizen	6,527	5,778	-749	-11%
18 years and over:	1,909,799	2,004,734	94,935	5%
Native	1,837,591	1,924,089	86,498	5%
Foreign born:	72,208	80,645	8,437	12%
Naturalized U.S. citizen	24,952	38,861	13,909	56%
Not a U.S. citizen	47,256	41,784	-5,472	-12%
Total Under 18	1,135,428	1,085,597	-49,831	-4%
Total Over 18	3,649,870	3,817,588	167,718	5%
CVAP	3,543,523	3,731,336	187,813	5%

Source: American Community Survey 2010-2019

**Appendix 4: 2010 – 2019 Black African American Estimated Population Changes in Alabama**

Label	B / AA 2010	B / AA 2019	# Difference	% Difference
Total:	1,262,980	1,319,551	56,571	4%
Male:	587,644	620,017	32,373	6%
Under 18 years:	178,038	166,682	-11,356	-6%
Native	177,700	166,122	-11,578	-7%
Foreign born:	338	560	222	66%
Naturalized U.S. citizen	45	216	171	380%
Not a U.S. citizen	293	344	51	17%
18 years and over:	409,606	453,335	43,729	11%
Native	404,676	447,911	43,235	11%
Foreign born:	4,930	5,424	494	10%
Naturalized U.S. citizen	1,819	3,829	2,010	111%
Not a U.S. citizen	3,111	1,595	-1,516	-49%
Female:	675,336	699,534	24,198	4%
Under 18 years:	171,215	152,776	-18,439	-11%
Native	171,003	151,923	-19,080	-11%
Foreign born:	212	853	641	302%
Naturalized U.S. citizen	96	313	217	226%
Not a U.S. citizen	116	540	424	366%
18 years and over:	504,121	546,758	42,637	8%
Native	498,700	541,156	42,456	9%
Foreign born:	5,421	5,602	181	3%
Naturalized U.S. citizen	2,208	3,916	1,708	77%
Not a U.S. citizen	3,213	1,686	-1,527	-48%
Total Under 18	349,253	319,458	-29,795	-9%
Total Over 18	913,727	1,000,093	86,366	9%
CVAP	907,403	996,812	89,409	10%

Source: American Community Survey 2010-2019

**Appendix 5: 2010 – 2019 Total Population Estimated Changes in Alabama**

Label	Hispanic 2010	Hispanic 2019	# Difference	% Difference
Total:	182,795	219,296	36,501	20%
Male:	98,816	109,914	11,098	11%
Under 18 years:	33,094	43,397	10,303	31%
Native	27,569	38,873	11,304	41%
Foreign born:	5,525	4,524	-1,001	-18%
Naturalized U.S. citizen	915	608	-307	-34%
Not a U.S. citizen	4,610	3,916	-694	-15%
18 years and over:	65,722	66,517	795	1%
Native	18,500	28,871	10,371	56%
Foreign born:	47,222	37,646	-9,576	-20%
Naturalized U.S. citizen	5,363	9,824	4,461	83%
Not a U.S. citizen	41,859	27,822	-14,037	-34%
Female:	83,979	109,382	25,403	30%
Under 18 years:	34,121	45,079	10,958	32%
Native	29,280	40,875	11,595	40%
Foreign born:	4,841	4,204	-637	-13%
Naturalized U.S. citizen	501	598	97	19%
Not a U.S. citizen	4,340	3,606	-734	-17%
18 years and over:	49,858	64,303	14,445	29%
Native	16,093	33,374	17,281	107%
Foreign born:	33,765	30,929	-2,836	-8%
Naturalized U.S. citizen	5,906	8,693	2,787	47%
Not a U.S. citizen	27,859	22,236	-5,623	-20%
Total Under 18	67,215	88,476	21,261	32%
Total Over 18	115,580	130,820	15,240	13%
CVAP	45,862	80,762	34,900	76%

Source: American Community Survey 2010-2019

## **Appendix 6: US Census**

### **Part 6a: What is the Census?**

The 2020 Census attempts to count every person living in the United States and the five U.S. territories. The goal is to count everyone only once and in the right place. The count is mandated by the Constitution and conducted by the U.S. Census Bureau. The requirement of taking a census is one of the first things mentioned in the U.S. Constitution, which provides some indication of how important a census was to the Founding Fathers (Voss and Cork, 2006).

The U.S. Constitution requires an “actual enumeration” of the population every 10 years in order to apportion seats in the House of Representatives among the states. States and localities also use census numbers for redistricting, to draw political boundary lines for their congressional delegations, legislatures, and other government districts. The census plays an important role in guiding the distribution of \$1.5 trillion in federal funding, as well as identifying needs for government services, such as schools and roads. Census statistics are the basis for a wide range of research and business decisions.

In a recent publication of the International Association of Official Statistics discussing the importance of Censuses in an international context, Everaers (2021) stated, “Population and Housing Censuses are an important cornerstone for National Statistical Systems. They provide a range of important statistics, relevant for policy-making, planning, and monitoring but also functioning as reference point and sample frame for many other national and regional statistics.” This description certainly applies to the U.S. Census. There is no single statistical resource more important than the Decennial Census.

In every census, the U.S. Census Bureau faces a trade-off between privacy protection and accuracy. According to the U.S. Census Bureau (2020d),

“One of the most important roles that national statistical offices (NSOs) play is to carry out a national population and housing census. In so doing, NSOs have two data stewardship mandates that can be in direct opposition. Good data stewardship involves both safeguarding the privacy of the respondents who have entrusted their information to the NSOs as well as disseminating accurate and useful census data to the public.”

The preceding suggests that this is an appropriate place to discuss privacy and confidentiality, two concepts that are often used interchangeably, but are distinct.<sup>15</sup> Privacy generally is used in regard to the right of an individual or organization to withhold information from others, while confidentiality is viewed as an extension of privacy in which an organization (such as the Census Bureau) that holds individual or organizational information is obligated to ensure that only authorized individuals have access to the information. While we will strive to maintain this distinction, the two concepts will inevitably overlap in this report.

---

<sup>15</sup> <https://research.uci.edu/compliance/human-research-protections/docs/privacy-confidentiality-hrp.pdf>

**Part 6b: Census Accuracy and Adjustments**

For over a century and for nearly as long as the Census Bureau has existed in its present form, it has had to balance its inherent, ingrained mission of collecting and producing high quality statistical information for the public good with a mandate to avoid disclosing information about any individual. In fact, the Census Bureau's mission is "to serve as the nation's leading provider of **quality data** about its people and economy." However, the mandate of "quality data" is tempered by an obligation to protect the privacy of Census respondents. The Census Bureau is bound by Title XIII of the United States Code. Title XIII provides the following protections to individuals and businesses:<sup>16</sup>

- Private information is never published. It is against the law to disclose or publish any private information that identifies an individual or business such, including names, addresses (including GPS coordinates), Social Security Numbers, and telephone numbers.
- The Census Bureau collects information to produce statistics. Personal information cannot be used against respondents by any government agency or court.
- Census Bureau employees are sworn to protect confidentiality. People sworn to uphold Title XIII are legally required to maintain the confidentiality of data. Every Census Bureau employee or contractor with access to personal data is sworn for life to protect your information and understands that the penalties for violating this law are applicable for a lifetime.

As part of this balancing act, the Census Bureau has used methods to help avoid disclosure of individual census respondents for many decades. According to the U.S. Census Bureau (2018), some method of disclosure avoidance has been used by the U.S. Census Bureau since 1970. However, as the privacy protections were put in place by the Census Bureau over the past several decades, there was never the threat of distorting the data as much as DP threatens to distort the 2020 Census data, and there was never the resistance seen among data users and demographers regarding the potential use of DP in the 2020 Census (Ruggles et al., 2019). The increase in resistance among data users reflects the extent to which they fear differential privacy will distort the data to the point that it is not usable for many functions.<sup>17</sup>

---

<sup>16</sup> [https://www.census.gov/history/www/reference/privacy\\_confidentiality/title\\_13\\_us\\_code.html](https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html)

**Part 6c: Census Bureau Privacy and Confidentiality and Title XIII**

Privacy, that is, the freedom to give or withhold information, and confidentiality, the government's obligations once it possesses the data, have been the most frequently raised concerns in the Twentieth Century with regard to the census. Privacy concerns and the public and private need for census information met head on in 1954 when Title XIII, the Census Act, was passed, which made responses to all census questionnaires mandatory. Title XIII U.S.C. §221, Chapter 7 states: "Whoever, being over eighteen years of age, refuses or willfully neglects, when requested by the Secretary ... to answer, to the best of his knowledge, any of the questions ... in connection with any census, shall be fined." Title 18 U.S.C. §3571 and §3559 provides that anyone over 18 years old who refuses or willfully neglects to answer questions posed by census takers of a fine of not more than \$5,000.

Even with Title XIII in place, privacy and confidentiality have been ongoing concerns with the census. It is important to note that the U.S. Census Bureau has used methods to help avoid disclosure of individual census respondents for many decades. According to U.S. Census Bureau (2018), some method of disclosure avoidance has been used since 1970 (Long, 2020). However, as the privacy protections were put in place by the Census Bureau, there was never the threat of distorting the data as much as DP threatens to distort the 2020 Census data.

**Part 6d: Uses of the Census**

To understand the importance of census accuracy, it is important to understand how census data are used. In addition to the scientific and scholarly interest in obtaining correct Decennial Census counts, there are many practical and policy-related reasons why it is important for Decennial Census data to be accurate. Census errors are important because they are both a data problem and, in many cases, a social equity issue.

Subnational Census inaccuracies are critical in terms of public policy consequences. The demographic numbers from the Census are used to distribute political power both in terms of assigning seats in Congress to states based on population and in the judicially mandated one-person/one-vote rule used for constructing political districts (Grofman 1982; McKay 1965; Balinski and Young 1982).

There is no definitive number of election districts where census data are used to draw boundaries for political districts. In addition to the 435 seats in Congress, almost all the 7,383 state legislators are elected from single member districts (National Conference of State Legislators 2017). Also, nearly every large city has council members elected from single-member districts, and the same is true for county commission seats in many jurisdictions. There are 19,355 elected county board members and elected executives, plus 18,629 independently elected officials result in 37,984 total county elected officials (including county board, executives and row officers).<sup>18</sup> School board members and many special districts also use census data to construct districts. Over 90,000 members make up the 49 U.S. States and the Virgin Islands School Board Associations.

State Decennial Census counts are used for apportioning the seats in the U.S. House of Representatives (Conk 1987), and sometimes small differences can be important. Crocker (2011) found that if 2010 Decennial Census count for North Carolina had been 15,753 higher it would have received an additional seat in Congress. This shows how small differences in counting might have large implications for political

---

<sup>18</sup> [https://www.naco.org/sites/default/files/documents/CM\\_2019.pdf](https://www.naco.org/sites/default/files/documents/CM_2019.pdf)

representation. Siegel (2002, Chapter 12) provides additional examples of how demographic data are used in a variety of political applications. The most recent estimates for Alabama show that a difference of as few as 5,000 people could make the difference between Alabama keeping its 7<sup>th</sup> congressional district or losing it to another state.<sup>19</sup>

Decennial Census data are also used in many federal funding formulas that distribute federal funds to states and localities each year (U.S. Senate 1992; Reamer 2009; Blumerman and Vidal, 2009). Recent research indicates there are 316 federal programs that use Census derived data to distribute more than \$1.5 trillion a year to states and localities (Reamer, 2019). The 55 largest federal programs that use census-derived data to distribute funds sent \$13.1 billion to Alabama in Fiscal Year 2016. The table below shows how much Alabama received from 16 large federal programs that use Census-derived data to distribute funds.

---

<sup>19</sup> [https://www.electiondataservices.com/wp-content/uploads/2020/12/NR\\_Appor20wTableMaps.pdf](https://www.electiondataservices.com/wp-content/uploads/2020/12/NR_Appor20wTableMaps.pdf)

**Table 6d1**

Selected Federal Assistance Programs That Distribute Funds on Basis of Decennial Census-Derived Data, U.S. and Alabama Fiscal Year 2016(2-28-2021)		
Program Name	Fiscal Year 2016 Obligations	
	U.S.	Alabama
Medical Assistance Program (Medicaid)	\$361,218,476,000	\$3,964,085,000
Supplemental Nutrition Assistance Program (SNAP)	\$66,376,250,674	\$1,254,835,320
Medicare Part B (Supplemental Medical Insurance) – Physicians Fee Schedule Services	\$66,076,784,523	\$1,129,410,997
Highway Planning and Construction	\$40,271,249,273	\$797,046,829
Section 8 Housing Choice Vouchers	\$19,387,184,000	\$194,272,000
Title I Grants to Local Education Agencies (LEAs)	\$14,364,454,918	\$230,728,658
National School Lunch Program	\$12,042,774,000	\$219,343,000
Special Education Grants (IDEA)	\$11,779,555,245	\$185,979,742
State Children's Health Insurance Program (S-CHIP)	\$13,761,924,000	\$457,272,000
Section 8 Housing Assistance Payments Program (Project-based)	\$10,156,542,138	\$105,166,471
Head Start/Early Head Start	\$8,648,933,810	\$138,342,659
Supplemental Nutrition Program for Women, Infants, and Children (WIC)	\$6,383,830,000	\$110,726,000
Foster Care (Title IV-E)	\$4,727,773,596	\$11,111,295
Health Center Program	\$4,319,604,643	\$76,252,531
Low Income Home Energy Assistance (LIHEAP)	\$3,351,810,105	\$43,520,240
Child Care and Development Fund – Entitlement	\$2,612,564,000	\$50,468,000
Total	\$645,479,710,925	\$8,968,560,742

Source: Reamer, 2017, Counting for Dollars, [https://gwipp.gwu.edu/sites/g/files/zaxdzs2181ff/downloads/IPP-1819-3%20CountingforDollars\\_AL.pdf](https://gwipp.gwu.edu/sites/g/files/zaxdzs2181ff/downloads/IPP-1819-3%20CountingforDollars_AL.pdf)

Demographic data are also used to distribute state government funds within states, but there is no good estimate of how much money is distributed by state governments based on census data (O'Hare 2020).

Many population projections also start with the Decennial Census counts, so differences such as those introduced by DP in the Decennial Census are likely to be reflected in projections for many years (U.S. Census Bureau 2014b). The 2010 Census figures are used as the base for the most recent Census Bureau (2014a) population estimates and projections (2014b). In discussing where to get data for state and local projections, Smith et al. (2001, page 113) indicate, "The most commonly used source--and the most comprehensive in terms of demographic and geographic detail--is the Decennial Census of population and housing."

In addition, Decennial Census results and the Census Bureau's post-census population estimates are often used to weight sample surveys both inside and outside government. If the Decennial Census counts and subsequent population estimates overestimate or underestimate a population group, the weighted survey results will reflect this error (Jensen and Hogan, 2017; O'Hare and Jensen 2014; O'Hare et al. 2013).

Data from the U.S. Decennial Census counts as well as projections which are based on the Census are used for many planning activities including schools (Edmonston 2001; McKibben 2007 and 2012). School systems in Alabama are examined more closely later in this report. In addition, data from the Census Bureau are often used as denominators for constructing rates such as the child mortality rates. Census



errors and data that have been significantly adjusted with methods such as DP may *significantly* skew such rates. These rates are based on using the Census counts as denominators.

For many groups, the Census is seen as a civil rights issue (Leadership Conference on Civil Rights 2017). In addition to heavy use of Decennial Census data in the context of voting rights, data from the Census are used to examine equality in jobs, housing, and education opportunities. A flawed census can undermine the ability to examine such issues. According to the Leadership Conference on Civil and Human Rights (2017, page 1), "Federal agencies rely on census and American Community Service (ACS) data to monitor discrimination and implement civil rights laws that protect voting rights, equal employment opportunity, and more."

Moreover, census inaccuracies can provide misleading public impressions about the size or growth of the population. This point is difficult to quantify, but in many instances the size of a population translates into the importance given to the population. In response to the 2000 Census, one public official stated, "Pride in the community is involved. I want people to really know how big we are. We aren't just a little burgh in south Louisiana." (cited in Prewitt 2003, page 7). It is difficult to overstate the places and ways in which census data are used, and the need for accuracy is critical.

The seminal document on issues with DP is 2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop (2020),<sup>20</sup> organized by the Committee on National Statistics (CNSTAT) of the National Academies of Sciences. Therein are numerous exceptional and detailed examples of the uses of small area Census Data and issues with the DP data produced to date. References will be made to proceedings of the workshop going forward in this document.

#### **School District Infrastructure**

One of the most important uses of accurate small-area data are school district infrastructure, enrollment, and forecasting. Small area Census data is used in the calculated population and enrollment forecasts for school districts at the district and attendance area level. These results are used to help districts with their short- and long-term planning on staffing, building utilization, and attendance area boundary modifications. Further, this research helps district balance their attendance area by size, socio-economic status, and race/ethnicity. Of particular importance is the age data that identifies the distribution of people throughout the course of life. Since school district attendance areas are not a census recognized "spine" geography, the data for these areas must be aggregated from block level Decennial Census data. The Decennial Census is the only source for reliable, valid, and accurate information of the demographic

---

<sup>20</sup> <https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations>

The Workshop on 2020 Census Data Products is organized by the Committee on National Statistics (CNSTAT) of the National Academies of Sciences, Engineering, and Medicine, at the request of the U.S. Census Bureau. The U.S. Census Bureau is implementing a new Disclosure Avoidance System (DAS) for the 2020 Census, after concluding that its previous methods permitted larger than expected risks of person reidentification. Implementing the 2020 DAS on 2010 Census data yielded a set of 2010 Demonstration Data Products, allowing data users to study impacts of the new system. The December 11-12, 2019, workshop provided a forum for studying the utility of privatized census tabulations and engaged the user and privacy communities in discussing trade-offs between "accuracy" and "privacy" in shaping the final 2020 DAS.

characteristics that can be used in this kind of research. Any additional, incremental error infused to the census results virtually eliminates the usefulness of this information. Given that there is no other source for this type of data at this level, its loss will be critical.

### **Consumer Demographics**

In the latter part of the 20th century, statistics became a commodity independent of government, and a statistical services industry developed. This development is pertinent because these services are primarily a business information industry. While demographics vendors such as Claritas and ESRI generate their own zip code estimates and forecasts, the Census Bureau uses both block and block group data to generate zip code population and housing estimates.<sup>21</sup> In the case of the zip code data generated by the Census Bureau, it is certain to be subject to DP if the latter is implemented; in the case of the zip code generated by demographic vendors, it is certain that the block and block group data they use in the process will be subject to DP if the latter is implemented. In 2018, for example, the Census Bureau decided to no longer approve requests for sub-state data if the data were not protected using strengthened disclosure avoidance methods providing small area data in its for-pay Custom Table operation.<sup>22</sup>

### **Health and Safety**

Small area data are important for public health and public safety, both for planning and reporting. As one example of the use of small area data for public health, the Thomas Jefferson Health District of Virginia (2016) developed a plan that incorporates census tract population data.<sup>23</sup> Similarly, in Clark County, Nevada, the Southern Nevada Healthy Food Access Program uses census tract population and income data.<sup>24</sup>

As another example, the US. Department of Energy issued a radiological monitoring plan<sup>25</sup> for the investigation of a proposed nuclear waste storage at Yucca Mountain, Nevada. The radiological studies area is defined by a circle 84 km in radius, whose center is assumed to be located at the proposed site of the central surface facilities (see Figure 1 below). The circle is divided into 160 cells radiated out from the 16 km radius area in the center of the study area, designated as the near field (NF) study area. The remainder of the area (16-84 km) is called the far field (FF) study area. The FF study area required that the population of each of the 160 cells be estimated on a regular basis (Swanson, Carlson, and Williams, 1990). As mentioned earlier, one researcher found the application of DP to the 2010 Census data would increase the teen pregnancy rate in one community from 5 percent to 66 percent. For someone concerned about adolescent health in that community, such a variation from the original census data would be very problematic.

---

<sup>21</sup> <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html>

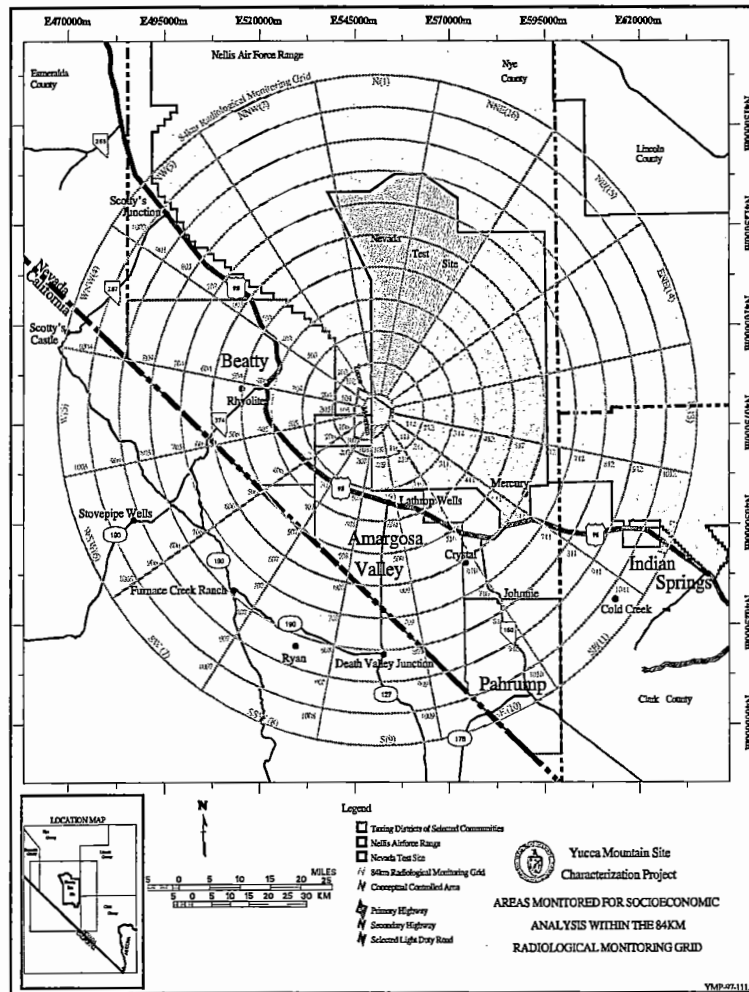
<sup>22</sup> <https://www.census.gov/programs-surveys/acs/data/custom-tables.html>

<sup>23</sup> <https://www.vdh.virginia.gov/content/uploads/sites/91/2016/07/2016-MAPP2Health-Report.pdf>

<sup>24</sup> <http://sns.rtcnv.com/wp-content/uploads/2019/11/Southern-Nevada-Healthy-Food-Access-Webmap-About-the-Data-11.2019.pdf>

<sup>25</sup> <https://www.nrc.gov/docs/ML0037/ML003753101.pdf>

Figure 6e1: The Radiological Studies Area.



**Natural Disaster Assessment**

Closely related to public health and safety, but distinct is natural disaster preparedness and assessment. As one example, Swanson and his colleagues (2009) examined the effect of Hurricane Katrina on the populations of 79 ZIP code areas in Louisiana (55) and Mississippi (24) devastated by the hurricane. Using the results by zip code, they estimated that Katrina reduced the area's overall population by 311,150 people (21.2%) from the 1,464,280 expected in the absence of Katrina.

In another study of the demographic effects of Hurricane Katrina, Swanson (2009) examined the effects of Hurricane Katrina on the client populations and candidates for a specific medical procedure in the service areas associated with two medical facilities on the Mississippi gulf coast. The two service areas were defined by zip codes, and in analyzing them, Swanson found that Katrina had an adverse impact on the client base of both medical facilities.

### **Regional Planning Organizations**

There are hundreds of regional planning organizations in the U.S.<sup>26</sup> Although they exist in every state, they may come under different names in different states, (Council of Government (COG), Metropolitan Planning Organizations (MPO)) but they all have similar missions, centered on land use and transportation planning, both of which require small area data.

TARCOG (Top of Alabama Regional Council of Governments) is an example of such an organization.<sup>27</sup> In its transportation planning, TARCOG uses block group data extensively (TARCOG, 2012). Like the plan issued by TARCOG, the 2015 transportation plan issued by the Montgomery, AL MPO issued makes use of small area data, including data representing census tracts.<sup>28</sup>

Although it is not a regional planning organization, Oak Ridge National Laboratory has developed LANDSCAN, a .geographically based population information system for the entire world.<sup>29</sup> The U.S. segment of the LANDSCAN system uses block data acquired from the Census Bureau (Bhaduri et al. 2007).

### **Public Use Microdata Samples (PUMS)**

The Public Use Microdata Samples (PUMS) are produced by the Census Bureau and primarily used by researchers. They are sets of individual and household records stripped of names and other information that could identify people. The Minnesota Population Center is perhaps the largest site in the U.S. where US and international PUMS (IPUMS), files can be accessed at no cost under the auspices of the "IPUMS" program.<sup>30</sup> The Minnesota Population Center is acutely concerned about DP and the error it will introduce into PUMS files (Ruggles et al., 2019).<sup>31</sup>

---

<sup>26</sup> [https://en.wikipedia.org/wiki/List\\_of\\_metropolitan\\_planning\\_organizations\\_in\\_the\\_United\\_States](https://en.wikipedia.org/wiki/List_of_metropolitan_planning_organizations_in_the_United_States)

<sup>27</sup> <http://tarcog.us/regional-planning-agency/>

<sup>28</sup> <http://montgomerympo.org/DOCS/2015/September23/Montgomery2040DraftLRTPAugust17.pdf>

<sup>29</sup> <https://landscan.ornl.gov/>

<sup>30</sup> <https://ipums.org/>

<sup>31</sup> <https://ipums.org/changes-to-census-bureau-data-products>

## **Section 7 Differential Privacy**

### **Part 7a: What is Differential Privacy?**

A statement by Ben Rossi (2016) is telling: "...[I]f a database is a representative sample of an underlying population, the goal of a privacy-preserving statistical database is to enable the user to learn properties of the population as a whole, while protecting the privacy of the individuals in the sample."

Since the application of differential privacy occurs within the Census Bureau's Disclosure Avoidance Systems (DAS), that term has sometimes been used to describe the use of differential privacy. To avoid confusion, the term differential privacy (DP) is used here to distinguish the version of DAS that includes DP from other versions of DAS.

This statement is telling because it reveals that the DP tradeoff is to make available properties of the population as a whole while protecting the privacy of individuals. In the world of the Census Bureau this tradeoff has been translated to mean that the population as a whole is defined by a population at a level of geography beyond the block. The tradeoff means that a user cannot learn properties of the population at the block level with any degree of confidence. If DP is implemented, it will affect all of the many users of small area data, to include those described earlier, the demographics vendors who supply clients with zip code level characteristics, public health and public safety organizations, and businesses that use small area data such as zip codes, school districts, and Regional Planning Organizations. The data associated with these census stakeholders are those that represent small areas directly as well as being aggregated into other small areas and into higher levels of geography. This means that DP, a statistical adjustment, will increase the error in the small area data needed by these stakeholders.

Is DP complicated? Here is a formal Definition followed by a discussion.<sup>32</sup>

**Definition 2.4 (Differential Privacy).** A randomized algorithm  $\mathcal{M}$  with domain  $\mathbb{N}^{|\mathcal{X}|}$  is  $(\epsilon, \delta)$ -differentially private if for all  $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$  and for all  $x, y \in \mathbb{N}^{|\mathcal{X}|}$  such that  $\|x - y\|_1 \leq 1$ :

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

Where,

M: Randomized algorithm i.e., query (db) + noise or query(db + noise).

S: All potential output of M that could be predicted.

x: Entries in the database. (i.e., N)

y: Entries in parallel database (i.e., N-1)

$\epsilon$ : epsilon, The maximum distance between a query on database (x) and the same query on database (y).

$\delta$ : Delta, the probability of information accidentally being leaked.

<sup>32</sup> Equation 1.1 in Lecture 1, Introduction to Differential Privacy: January 28, CSE711, Topics in Differential Privacy, SUNY Buffalo, Spring, 2016, M. Gaboardi  
<https://www.acsu.buffalo.edu/~gaboardi/teaching/cse711Spring2016/Lecture1.pdf>.

This definition of DP is a measure of “How much privacy is afforded by a query?” This is an important point in that DP represents an offer of privacy according to a provable and quantifiable amount, sometimes referred to as the privacy-loss budget (Snoke and McKay, 2019). It is this probabilistic quantifiable feature that is DP’s major selling point because other forms of DAS (Disclosure Avoidance Systems) do not provide a formal quantification of the protection they offer. How does it do this? As this suggests, DP is not the system that creates privacy; it is the system that measures privacy using the definition just given. How does DP measure privacy?

The DP algorithm gives the comparison between running a query  $M$  on database  $(x)$  and on a parallel database  $(y)$ , where the latter has one less entry than database  $(x)$ . The measure by which the full database  $(x)$  and the parallel database  $(y)$  can differ is given by Epsilon ( $\epsilon$ ) and delta ( $\delta$ ). Specifically, DP works by tying privacy to how much the answer to a question or statistic is changed given the absence or presence of the most extreme possible person in the population. This is done within a statistical framework. An example by Snoke and McKay (2019) helps to explain this. Suppose the data we want to protect is income data, and the statistic we want answered is, “What is the median income?” The most extreme person who could possibly be in any given income data could be Jeff Bezos. If he is absent or present in the data set, the median will not change much, if at all. This means that DP can provide a more accurate answer about the median income without using much privacy-loss budget.

However, what if the question is, “What is the maximum income?” Unlike the median, the answer to this question would be likely to significantly change if Bezos is absent or present in the data set. A DP algorithm would provide a less accurate answer, or require more privacy-loss budget, to answer this query and protect the extreme case, Bezos (Snoke and McKay, 2019).

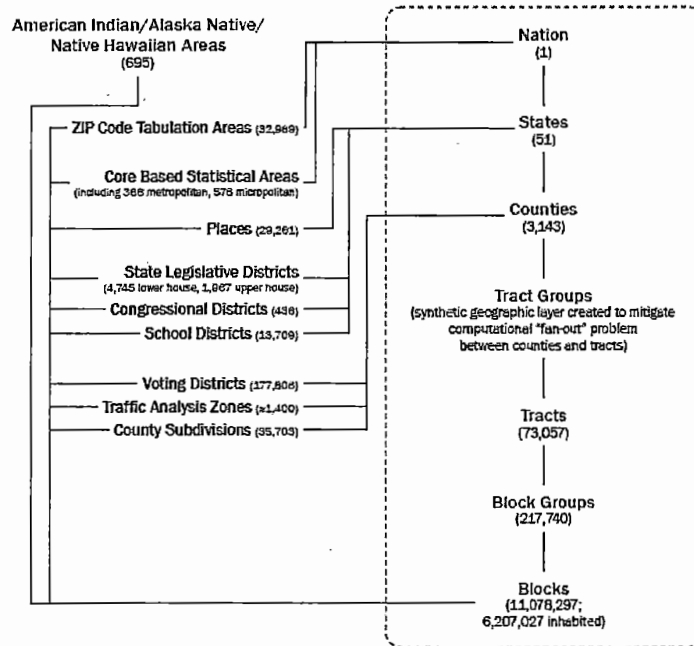
So, when Epsilon ( $\epsilon$ ) is small, DP asserts that for all pairs of adjacent databases  $x, y$  and all outputs  $M$ , an adversary cannot distinguish which is the true database on the basis of observing the output—the probabilities are too low. That is, if we are interested in median income, it does not matter if Jeff Bezos is in or out of the data set: For this query Epsilon ( $\epsilon$ ) should be set at a high level because for a query regarding median income there is little need to “protect” the data base. This example translates formally into something like the following. When ( $\epsilon$ ) is large DP merely says that there exists neighboring databases and an output  $M$ , for which the ratio of probabilities of observing  $M$  conditioned on the database being, respectively,  $x$  or  $y$ , is large.

However, if we are interested in knowing the maximum income in the data base, it will matter if Jeff Bezos is in or out of the database. Thus, Epsilon ( $\epsilon$ ) should be set at a low value in order to prevent “leaking” the maximum income. However, even if Epsilon ( $\epsilon$ ) is not set low, an adversary may not have the right auxiliary information to recognize that a revealing output has occurred; or may not know enough about the database(s) to determine the value of their difference.

As you can see, the DP algorithm represents a statistical adjustment in that it uses a probability framework, typically based on the Laplace probability distribution (as stated elsewhere in this report), which is used to produce the errors / noise in the data. Moreover, as noted by Ruggles et al. (2019) under DP, responses of individuals cannot be divulged even if the identity of those individuals is unknown and cannot be determined. Returning to the example of a query about maximum income, it would not matter if the identify of Bezos was not divulged; the correct answer to the question about the maximum income in a dataset would not be provided under DP.

A final important point about differential privacy is how it is applied geographically. In our analysis, we look at two different types of geography: “spine” which are the core census statistical geographies such as counties, tracts, and blocks, and “off-spine” which are governmental or administrative geographies such as school districts and legislative districts. The “spine” geography, particularly blocks, are important because they offer the greatest geographic granularity and are the geographies DP is actually being applied to. “Off-spine” geographies are also critically important because conceptually they could capture the best or worst pieces of statistical geography and aggregate and magnify their errors. As shown in Figure X.X (above), legislative districts, voting districts, congressional districts, places, VTDs, and ZIP codes are all “off-spine,” that is, not in the hierarchy of geographic areas for which the TopDown Algorithm (TDA) maximizes accuracy and so are built up from the lower-level block groups and blocks.<sup>33</sup>

**Figure 7a1: Hierarchy of census geographic entities, with reference to generation of the 2010 Demonstration Data Products.**



**Source:** 2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop Page 33.

<sup>33</sup> 2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop Page 67

**Part 7b: How is Differential Privacy being Proposed to be Used in the 2020 Census**

Historically, surveys and the census require respondents to reveal sensitive information under the promise that such information will remain confidential. Traditionally, protection from disclosure was accomplished by anonymizing records. In this way,

“statistical analyses on issues of public importance could be accomplished while protecting the identity of the respondent. Over time however, the availability of public external data and the increase in capability of data analytics has made protecting confidential data a challenge. By linking information in one data set with that of another containing some intersecting information (known as a record-linkage attack) it is sometimes possible to connect an anonymous record containing confidential information with a public record and thus identify the respondent. This is called re-identification of previously de-identified data.” (Long, 2020)

This is exactly the kind of reidentification the Census Bureau is trying to protect against. While there have been a number of newsworthy reidentifications, there have been no known cases against the 2010 Census. (Ruggles et al., 2019). After several years of developing the method and infrastructure, in October 2019, the Census Bureau released a demonstration data product to help users assess the impact of differential privacy on the utility and accuracy of Decennial Census data. This product was a differentially private version of the 2010 Decennial Census. Several assessments of the demonstration data were presented at the Workshop on 2020 Data Products (December 11-12, 2019) organized by the Committee on National Statistics of the National Academy of Sciences. These assessments identified limitations in the differentially private data, particularly for low-population geographic units, for which there are no other sources of complete, reliable population data. Workshop participants urged the Bureau to release additional demonstration data as they work to improve utility by refining the differentially private algorithm.<sup>34</sup>

The problem that DP is designed to fix is complicated, as is the implementation of DP. The passage below from the U.S. General Accountability Office (2020, page 14) is a good, short description of this issue.

“Differential privacy is a disclosure avoidance technique aimed at limiting statistical disclosure and controlling privacy risk. According to the Bureau, differential privacy provides a way for the Bureau to quantify the level of acceptable privacy risk and mitigate the risk that individuals can be reidentified using the Bureau’s data. Reidentification can occur when public data are linked to other external data sources. According to the Bureau, using differential privacy means that publicly available data will include some statistical distortions, noise, or data inaccuracies, to protect the privacy of individuals. Differential privacy provides algorithms that allow policy makers to decide the trade-offs between data accuracy and privacy.”

---

<sup>34</sup> Source: <https://www.nhgis.org/privacy-protected-demonstration-data#purpose>

Reference: <https://www.census2020now.org/challenges-blog/2019dp>

Reference: [https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html#par\\_list](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html#par_list)



Basically, DP injects intentional error into the census tabulations that are based on the true responses to the census by adding or subtracting random numbers from table cells that reflect the true responses. Adding or subtracting random numbers to the census is intended to make it more difficult to specific respondents. While the process of introducing numbers may be random, it is important to note that the outcomes are anything but. As seen in the analytics section, errors are much more heavily concentrated in NVA children, minorities, and off-spine geographies. The U.S. Census Bureau (2020e) provides more information on the use of DP in the 2020 Census along with regular updates of their work (U.S. Census Bureau 2020c). For an independent look at differential privacy see Boyd (2020).

#### **Part 7c: Differential Privacy and the Census: Existing Concerns from the User Community**

Many researchers, demographers, and data users have expressed concerns about the possible use of DP in the 2020 Census. A few of the comments expressing worries about the use of DP are provided below.

The National Academy of Sciences, Committee on National Statistics (CNSTAT) Workshop held December 11-12, 2019, titled “Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations” provides a lot of data related to the accuracy of the Census Bureau’s October 2019 Demonstration Product (Committee on National Statistics 2019). Many of the presenters at this conference expressed apprehensions about the potential use of DP in the 2020 Census.

Based on the evidence presented at the CNSTAT workshop and their own internal analysis, the U.S. Census Bureau (2020b) concluded, “The October Vintage of the DAS falls short of ensuring ‘Fitness for use’ for several priority use cases.” This led to subsequent versions of DP-infused data being released by the Census Bureau.

Two prominent demographers Hotz and Salvo (2020) helped organize the December 2019 CNSTAT conference and later provided a review and summary of the event. They concluded,

“At the same time, evidence presented at the Workshop indicated that most data for small geographic areas – especially census blocks – are not usable given the privacy-loss level used to produce the demonstration file.”

In their summary Hotz and Salvo go on to say, “Many within the community remain skeptical of the Bureau’s adoption of differential privacy and its consequences for the use cases.”

Many presenters at the Committee on National Statistics Conference found impossible or improbable results when DP was applied to the 2020 Census data. For example, with respect to Differential Privacy (DP) the state demographer of Virginia (Qian Cai, 2020) noted,

“As another example Cai mentioned that DP information for the small city of Emporia suggests that the city’s teen pregnancy rate would shift from 5 percent to 66 percent, which has definite implications. “

At a recent research conference, Cropper and Stojakovic (2021) examined the use of DP for data often used in the context of school enrollment projections: namely population pyramids which is the population broken down by age and sex. Population pyramids are used to calculate the expected number of future births based on the age and sex composition of the population in a school district. In discussing the implications of DP for school district demography, they conclude: “Although total population at the school

district level is relatively close pre- and post-DAS, the error in age/sex cohorts is very problematic.” They go on to say, “Working with Census 2020 population less than 100,000 (in terms of age/sex) may not even be reliable/usable.”

The impact of DP on census data may have negative implications for health matters. In a very recent publication, after examining the impact differential privacy has on Covid-19 Rates, Hauer and Santos-Lozada (2021, page 1) conclude,

“Using empirical COVID-19 mortality curves, the authors show that differential privacy will introduce substantial distortion in COVID-19 mortality rates, sometimes causing mortality rates to exceed 100 percent, hindering our ability to understand the pandemic. This distortion is particularly large for population groupings with fewer than 1,000 persons: 30 percent of all county-level age-sex groupings and 60 percent of race groupings.”

At a research conference held in February 2021, Dr. Richelle Winkler and several colleagues examined the implications of DP for a widely used product based on Decennial Census data. The product is county-specific net migration rates that have been produced from Decennial Census data for the past several decades. They concluded that “inaccuracies in DP data pose critical challenges to NME accuracy.” (NME is Net Migration Estimates.)

Researchers are not the only ones raising concerns about differential privacy. In a letter dated September 21, 2020, 33 members of Congress sent a letter to the Census Bureau director raising some concerns about differential privacy. They said, “We write to express concern with the U.S. Census Bureau’s proposed “differential privacy” approach to maintain the confidentiality of data collected in the 2020 Decennial Census.” Near the end of the letter, they say, “As Members of Congress, we are concerned about the potential unintended consequences that a differential privacy method could have on the allocation of important funding and other activities that rely on census data.”

Redistricting is one the most important uses of census data, and jeopardizing accuracy by using DP could result in unfair districts being constructed. Several organizations have raised concerns about what the use of DP might do to the redistricting process.

In a letter to the Census Bureau dated May 14, 2020, The National Conference of State Legislatures stated, “The Census Bureau’s decision to use differential privacy as its statistical method to meet the goal of avoiding the disclosure of individual response may not be the best method to ensure states receive the most accurate data for redistricting purposes.”

In a letter to the Census Bureau dated February 13, 2020, the Utah State Legislature stated, “Based upon our analysis of differential privacy as applied to the 2010 Census redistricting data, we believe, if differential privacy is applied to the 2020 redistricting data, that the integrity of the data used to redistrict the state into congressional and legislative districts, and also with local jurisdictions will be threatened.”

In a letter to the Census Bureau dated April 24, 2020, the head of National Redistricting Foundation stated, “I write today on behalf of the National Redistricting Foundation (“NRF”) to convey our significant concerns regarding the Census Bureau’s proposed use of differential privacy for the 2020 Census. We are concerned that the Bureau’s proposed application of differential privacy will substantially diminish the

usability of the resulting data for redistricting, hampering the ability of state and local governments to comply with constitutional and statutory requirements that ensure fair and equal political representation.”

In a report from the U.S. General Accountability Office in December 2020 related to the 2020 Census, they label the last section in the report “The Bureau Has Work Remaining to Protect the Privacy of Respondents’ data.” In this section they note several things that the Census Bureau must do before implementing differential privacy.

In another significant example, Toni G. Atkins (California Senate President pro Tempore) and Anthony Resdon (California Speaker of the Assembly) have written to the Honorable Ronald Klain (the current Presidential Chief of Staff), copying Vice President Kamala Harris (among others), stating:

“The adjustment to Census data that remain a concern is referred to as “differential privacy.” It has been developed by the Census Bureau as part of its mandate to maintain the confidentiality of individual American residents, while at the same time producing accurate detailed data. The Census Bureau is required to balance these two goals – which can be in conflict.

Differential privacy is a new system that the previous administration rushed to complete to avoid disclosing individuals’ identities. The intent was to serve the laudable privacy goal, but the system also has the effect of scattering minority voters, making it much more difficult to serve the goals of the Voting Rights Act.

The rush to implement differential privacy also negatively impacts the ability of states to implement their laws. For example, California requires Census data to be adjusted for persons incarcerated in state correctional facilities, using the data provided by the Census Bureau. However, if this data has been modified by the Bureau with limited transparency, California’s important electoral reform will be undermined. We include an enclosure describing this particular problem in the California setting in more detail, but the issue of this new adjustment’s impact is a national one.”<sup>35</sup>

Ruggles et al. (2019: 404) observe that the application of differential privacy to census data products is a radical departure from established Census Bureau confidentiality laws. They go on to note that differential privacy requirement that database outputs do not significantly change when any individual data are added or removed has implications, especially the aspect under differential privacy in which it is prohibited to reveal characteristics of an individual even if the identity of that individual is effectively concealed. Continuing, Ruggles et al. (2019: 404) point out that as the Census Bureau acknowledges, masking respondent characteristics is not required under census law. Instead, the laws require that the identity of particular respondents shall not be disclosed. In 2002, Congress explicitly defined the concept of identifiable data: “It is prohibited to publish any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.”

---

35

[https://www.ncsl.org/Portals/1/Documents/Redistricting/California\\_Leaders\\_Letter\\_to\\_RonaldKlain\\_Feb2021.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/California_Leaders_Letter_to_RonaldKlain_Feb2021.pdf)

## References

- Abowd, J (2020). Modernizing Disclosure Avoidance: What We've Learned, Where We Are Now. [https://www.census.gov/newsroom/blogs/research-matters/2020/03/modernizing\\_disclosure.html](https://www.census.gov/newsroom/blogs/research-matters/2020/03/modernizing_disclosure.html)
- Anderson, M. and Fienberg, S.E., (2001). Who Counts? The Politics of Census Taking in Contemporary America, Russell Sage Foundation, New York.
- Balinski, M. and Young, H. P., (1982). Fair Representation: Meeting the Ideal of One Man, One Vote, Yale University Press, New Haven, CT.
- Blumerman, L.M., and Vidal, P.M., (2009). Uses of Population and Income Statistics in Federal Funds Distribution—With a Focus on Census Bureau Data, Government Divisions Report Series, Research Report #2009-1, U.S. Census Bureau, Washington, DC.
- Brace, K. (2016) "No Change in Apportionment Allocations with New 2016 Census Estimates: But Greater Changes Likely by 2020." Election Data Services, December
- Bryant, B.E., Dunn, W. (1995). Moving Power and Money: The Politics of U.S. Decennial Census Taking, New Strategists Publications, Ithaca NY.
- Calleam Consulting (2012). Failed Project Case Study -The United States 2010 Census – Field Data Collection Automation (FDCA). ([http://calleam.com/wp-content/uploads/US\\_Census\\_FDCA\\_Case\\_Study\\_V1.0.pdf](http://calleam.com/wp-content/uploads/US_Census_FDCA_Case_Study_V1.0.pdf))
- Clogg, C.C., Massaglie, M.P, and Eliason, S.R., (1989). Population Undercount and Social Science Research, Social Indicators Research, Vol. 21, no. 6, pp 559-598.
- Committee on National Statistics (2019). "Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations," presentations are available at <https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations>
- Conk, M., (1987). According to Their Respective Numbers, Yale University Press, New Haven, CT.
- Conk, M., (1987). According to Their Respective Numbers, Yale University Press, New Haven, CT.
- Crocker, R. (2011). House Apportionment 2010: States Gaining, Losing and on the Margin, Congressional Research Service, 7-5700 R41584.
- Cropper, M. and Stojakovic, Z (2021) "The Impact of Differential Privacy on School Demography and Small Area Forecasts," presentation at the Applied Demography Conference February 2-4, 2021, Final slide in presentation.
- Daponte, B.O., and Wolfson, L.J., (2003). How many American Children are Poor? Considering Census Undercounts by Comparing Census to Administrative Data," unpublished paper.
- Dinur, I. and K. Nissim. (2003). Revealing Information while Preserving Privacy. In PODS, pages 202–210. ACM.
- Dwork, C. (2006) Differential Privacy (<https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dwork.pdf>).

Edmonston, B., (2001). Effects of U.S. Decennial Census Undercoverage on Analyses of School Enrollments: A Case Study of Portland Public Schools, U.S. Census Monitoring Board, Report Series, Report No. 5, February

Everaers, P. (2021) Editorial, Statistical Journal of the International Association of Official Statistics: Statistical Journal of the IAOS, vol. 36, no. 1, pp. 1-3, 2020 DOI-10.3233/SJI-209002.

First Focus on Children (2019) Children's Budget 2019

Grofman, B., (1982). Representation and Redistricting Issues, Lexington Books, Lexington, MA.

Hauer, M. E. and A. Santos-Lozada. (2021) "Differential Privacy in the 2020 Census Will Distort COVID-19 Rates, Socius: Sociological Research for a Dynamic World,

Hernandez, D. and N. Denton. (2001). Census Affects Children in Poverty, U.S. Census Monitoring Board, Washington, DC.

Hough, G., and Swanson, D. (2006). An evaluation of the American Community Survey: Results from the Oregon test site. Population Review and Policy Research 25(3): 257-273.

Hotz, J. and Salvo J. (2020). Addressing the Use of Differential Privacy for the 2020 Census: Summary of What We Learned from the CNSTAT Workshop. <https://www.apdu.org/2020/02/28/apdu-member-post-assessing-the-use-of-differential-privacy-for-the-2020-census-summary-of-what-we-learned-from-the-cnstat-workshop/>.

Jensen, E., and H. Hogan (2017). The coverage of young children in demographic surveys. Statistical Journal of the International Association for Official Statistics 33: 321-333.

Leadership Conference on Civil and Human Rights (2017). Fact Sheet: The Census and Civil Rights," Downloaded on June 13, 2017 from <http://civilrightsdocs.info/pdf/census/Fact-Sheet-Census-and-Civil-Rights.pdf>

Long, G., (2020) Formal Privacy Methods for the 2020 Census. Mitre.Org

McKay, R., (1965). Reapportionment: The Law and Politics of Equal Representation, The Twentieth Century Fund, New York, NY.

McKibben, J., (2007). The Use of School Enrollment Data to Estimate Census Undercounts in Small Areas, presentation and Applied Demography Conference, San Antonio, TX January 2007.

McKibben, J., (2012). Using School Enrollment Data to Measure Small Area Coverage Rates of the 2010 Census, presentation and Applied Demography Conference, San Antonio, TX January 2012.

Murphy, S.L. Xu, J., Kochanek, K.D., (2013). Deaths: Final data for 2010, National Vital Statistics Reports, Volume 61, No. 4.

Nagle, N. and Kuhn, T. (2019). "Implications for School Enrollment Statistics." <https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations>

National Conference of State Legislators (2017) downloaded June 11, 2017 from <http://www.ncsl.org/research/about-state-legislatures/number-of-legislators-and-length-of-terms.aspx>

National Conference of State Legislators (2017) downloaded June 11, 2017 from <http://www.ncsl.org/research/about-state-legislatures/number-of-legislators-and-length-of-terms.aspx>

O'Hare, W.P., Jensen, E. and O'Hare, B.C., (2013). Assessing the Undercount of Young Children in the U.S. Decennial Census: Implications for Survey Research and Potential Explanations. Paper presented at the 2013 American Association of Public Opinion Researchers Annual Conference, Boston, MA.

O'Hare, W.P., (2014c). State-Level 2010 Census Coverage Rates for Young Children, Population Research and Policy Review, Volume 33, no. 6, pages 797-816.

O'Hare, W. P., and Jensen, E. B., (2014). The Representation of Young Children in the American Community Survey, presentation at the ACS Users Group Conference, Washington, DC. May 29-30.

O'Hare, W.P. (2017) presentation at American Community Survey User Conference, 2017, Alexandria, VA

O'Hare W. P. (2020). "Many States Use Decennial Census Data to Distribute State Money, The Census Project Website <https://thecensusproject.org/2020/01/09/many-states-use-decennial-census-data-to-distribute-state-money/>

O'Hare, W.P. (2020). "The Politicization of the 2020 Census," PAA Affairs, Fall 2020, The Population Association of America, Washington DC. [https://higherlogicdownload.s3.amazonaws.com/POPULATIONASSOCIATION/3e04a602-09fe-49d8-93e4-1dd0069a7f14/UploadedImages/Documents/PAA\\_Affairs/PAA-Fall\\_20\\_.pdf](https://higherlogicdownload.s3.amazonaws.com/POPULATIONASSOCIATION/3e04a602-09fe-49d8-93e4-1dd0069a7f14/UploadedImages/Documents/PAA_Affairs/PAA-Fall_20_.pdf)

O'Hare, W.P, Jacobsen, L.A., Mather, M. VanOrman and Pollard, K (2019). "What Factors Are Most Closely Associated with the Net Undercount of Young Children in the U.S. Census?" Population Reference Bureau, Washington, DC. <https://www.prb.org/wp-content/uploads/2019/03/net-undercount-children-acs.pdf>

O'Hare, W.P. (2019). "Evidence mounts regarding respondent confusion about counting young children in the Census" The Census Project <https://thecensusproject.org/2019/12/11/evidence-mounts-regarding-respondent-confusion-about-counting-young-children-in-the-census/>

Prewitt, K. (2003). Politics and Science in Census Taking, in series The American People: Census 2000, Russell Sage Foundation and Population Reference Bureau, Population Reference Bureau, Washington, DC.

Prewitt, K., (2014). What Is Your Race? The Census and Our Flawed Efforts to Classify Americans, Princeton University Press, Princeton, NJ.

Price Waterhouse Cooper (2001). Effect of U.S. Decennial Census 2000 Undercount on Federal Funding to States and Selected Counties, 2001-2012, Report to the U.S. Census Monitoring Board, Presidential Members.

Qian, C., (2020) Census Data Products, Data Needs and Privacy Considerations, Proceedings of a Workshop, The National Academy Press, page 28, Washington DC.

Reamer, A. D., (2009). Counting for Dollars, Brookings Institute, Metropolitan Policy Program, Washington, DC.

Reamer, A. D., (2019). Counting for Dollars, George Washington University, Washington, DC.

Reamer, A. D., (2020). Counting for Dollars, George Washington University, Washington, DC.

Ruggles, S., C. Fitch, D. Magnuson, and J. Schroeder. (2019) Differential Privacy: Implications for Social and Economic Research. American Economic Association Papers and Proceedings 109 (May): 403-408.

Samuels, C.A. (2017) "Preschool Class Size – Within Reason – Does It Matter Study Finds" Education Week, August 4 2017.

Schwede, L., Terry, R., and Hunter, J. (2015). "Ethnographic evaluations on coverage of hard-to-count minority in the US decennial censuses," in Hart-to-Survey Populations, Edited by Tourangeau, R., Edwards, B., Johnson, T.P., Wolter, K.M., and Bates, N. Cambridge University Press, Cambridge, England, pp 293-315

Smith, S., Tayman, J., and Swanson, D.A., (2001) State and Local Population Projections: Methods and Analysis, The Plenum Series on Demographic Methods and Population Analysis, Kluwer

Snoke, J., and C. McKay. (2019). Differential Privacy: What is it? AMSTAT News (March). (<https://magazine.amstat.org/blog/2019/03/01/differentialprivacy/>).

Swanson, D. (2009). "Hurricane Katrina: A Case Study of Its Impacts on Medical Service Providers and Their Client Populations." The Open Demography Journal 2: 8-17.

Swanson, D., J. Carson, and C. Williams. (1990). "The Development of Small Area Socioeconomic Data to be Utilized for Impact Analysis: Rural Southern Nevada." pp.985-990 in High Level Radioactive Waste Management: Proceedings of the 1990 International Conference, American Nuclear Society and American Society of Civil Engineers, New York, New York, 1990.

Swanson, D., R. Forgette, J. McKibben, M. Van Boening, and L. Wombold. (2009). "The Socio-Demographic and Environmental Effects of Katrina: An Impact Analysis Perspective". The Open Demography Journal.2 (11): 36-46.

The Leadership Conference Education Fund/Advancing Justice/National Association of Latino Elected Officials (2014). Race and Ethnicity in the 2020 Census: Improving Data to Capture a Multiethnic America, Leadership Conference Education Fund, Washington DC.

U.S. Senate (1992). Dividing Dollars: Issues in Adjusting Decennial Counts and Intercensal Estimates for Funds Distribution, Report prepared by the Subcommittee on Government Information and Regulation of the Committee on Government Affairs, 102nd Congress, 2nd session Senate Print 102-83, U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (1974). Estimates of Coverage of Population by Sex, Race and

U.S. Census Bureau (2014a) (BILL UPDATE). The 2013 population estimates are available online at <http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk>

U.S. Census Bureau (2014b). U.S. Population Projections: 2014-2060, Release Number CB14-TPS.86.

U.S. Census Bureau (2017) 2020 Census Operation Plan, Version 3.0, Issued September 2017, U.S. Census Bureau, Washington, DC.

U.S. Census Bureau (2018), "Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing," THE RESEARCH AND METHODOLOGY DIRECTORATE, Mc Kenna, L. U.S. Census Bureau, Washington DC., <https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf>

U.S. Census Bureau (2020a). 2020 Census Disclosure Avoidance Improvement Metrics, U.S. Census Bureau, Washington DC., March 18, <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf?#>

U.S. Census Bureau (2020b), "2020 Census Data Products and the Disclosure Avoidance System, Hawes M. and Garfinkel. S. L., Planned presentation at the Census Scientific Advisory Committee meeting, March 26,

U.S. Census Bureau (2020c) DAS Updates, U.S. Census Bureau, Hawes M. June 1 Washington DC., <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-06-01-das-updates.pdf?#>

U.S. Census Bureau (2020d). "Disclosure Avoidance and the Census," Select Topics in International Censuses, U.S. Census Bureau, October 2020. <https://www.census.gov/library/working-papers/2020/demo/disclos-avoid-census.html>

U.S. Census Bureau (2020e). "Disclosure Avoidance and the 2020 Census, U.S. Census Bureau," Washington DC., Accessed November 2, <https://www.census.gov/about/policies/privacy/statistical-safeguards/disclosure-avoidance-2020-census.html>

U.S. Census Bureau (2020f) Error Discovered in PPM, U.S. Census Bureau, Washington DC. <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>

U.S. Census Bureau (2020g), "2020 Disclosure Avoidance System Updates," U.S. Census Bureau, Washington DC., <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>

U.S. General Accountability Office (2020). "COVID-19 Presents Delays and Risks to Census Counts," U.S. General Accountability Office, Washington, DC., <https://www.gao.gov/products/GAO-20-551R>

U.S. Senate (1992). Dividing Dollars: Issues in Adjusting Decennial Counts and Intercensal Estimates for Funds Distribution, Report prepared by the Subcommittee on Government Information and Regulation of the Committee on Government Affairs, 102nd Congress, 2nd session Senate Print 102-83, U.S. Government Printing Office, Washington, DC.

Vink, J. (2019). "Elementary School Enrollment," <https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations>

Voss, P. and D. Cork (2006). Once, Only Once, and in the Right Place: Residence Rules in the Decennial Census. National Research Council. Washington, DC: The National Academies Press.

Waite, P. J. (2003). The reengineered 2010 Census. Proceedings of Statistics Canada Symposium 2003 Challenges in Survey Taking for the Next Decade (<https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X20030017608>).

Waite, P. J. and B Reist (2005). Reengineering the census of population and housing in the United States Statistical Journal of the United Nations Economic Commission for Europe, 22 (1): 13-23



Weldon Cooper Center for Public Service (2013) Projections for the 50 States and D.C., available online at <http://www.coopercenter.org/demographics/national-population-projections>

West, K.K, and Fein, D.J., (1990). U.S. Decennial Census Undercount: An Historical and Contemporary Sociological Issues, Sociological Inquiry, Vol. 60, No. 2, May pp 127-141.

Williams, J. (2012). The 2010 Census: Background and Issues. Congressional Research Service. (<https://www.census.gov/history/pdf/2010-background-crs.pdf> ).

Winkler, R. L., Butler, J.L, Curtis, K. J., Robinson, D. E. (2021) "Impact of Differential Privacy on New Net Migration Estimates (2010-2020) presentation at the PAA Applied Demography Conference, February 2-4, 2021.

## **Appendix 2**

## Thomas M. Bryan

3132 Briarmoor Lane

Midlothian, VA 23113

425-466-9749

tom@bryangeodemo.com

### Résumé and C.V.

#### Introduction

I am an applied demographic, analytic and research professional. I have expertise in the collection, management, analysis and reporting of demographic, business and consumer data to deliver insights and drive decision making. I have subject matter expertise in:

- Political redistricting and Voting Rights Act related litigation
- US Census Bureau data and national health statistics
- Large-scale multi-mode consumer survey research design and execution
- Applied demographic techniques
- Advanced analytics
- Consumer package goods market information and consumer research
- FDA compliance and the Family Smoking Prevention and Tobacco Control Act
- Geographic Information Systems (G.I.S.)
- U.S. Government, Census and other primary / secondary survey research data (NHIS, BRFSS, NSDUH)
- Syndicated data and vendor management (IRI, Nielsen, GfK, ORC Engine, etc.)

#### Education & Academic Honors

2002 MS, Management and Information Systems - George Washington University

2002 GSA CIO University graduate\* - George Washington University

1997 Graduate credit courses taken at University of Nevada at Las Vegas

1996 MUS (Master of Urban Studies) Demography and Statistics core - Portland State University

1996 Oregon Laurels Scholar

1992 BS, History - Portland State University

1987-1988 Undergraduate credit courses taken at Portland OR, Community College

---

Granted by the General Services Administration (GSA) and the Federal IT Workforce Committee of the CIO Council.

<http://www.gwu.edu/~mastergw/programs/mis/pr.html>

**Bryan GeoDemographics, January 2001-Current: Founder and Principal**

I founded Bryan GeoDemographics (BGD) in 2001 as a demographic and analytic consultancy to meet the expanding demand for advanced analytic expertise in applied demographic research and analysis. Since then, my consultancy has broadened to include litigation support, political redistricting, school redistricting, and municipal infrastructure initiatives. Since 2001, BGD has undertaken over 150 such engagements in three broad areas: 1) state and local political redistricting, 2) applied demographic studies, and 3) school redistricting and municipal Infrastructure analysis.

The core of the BGD consultancy has been in political redistricting litigation, particularly on Voting Rights Act and discrimination cases. Engagements include:

State and Local Political Redistricting

- 2020: In the matter of *The Christian Ministerial Alliance (CMA), Arkansas Community Institute, Marion Humphrey, Olly Neal, And Ryan Davis v. the State of Arkansas*. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Providing demographic and analytic litigation support.
  - [https://www.naacpldf.org/wp-content/uploads/CMA-v.-Arkansas FILED-without-stamp.pdf](https://www.naacpldf.org/wp-content/uploads/CMA-v.-Arkansas%20FILED-without-stamp.pdf)
- 2020: In the matter of *Louisiana State Conference of the NAACP, Allen and Anthony (Plaintiffs) v. the State of Louisiana (Defendants)* in US District Court. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Providing demographic and analytic litigation support for the analysis and testing of LA Supreme Court Districts.
  - <https://apnews.com/c44c986a29ec4035a87e5ca94d4e6324>
  - <https://www.bloomberglaw.com/public/desktop/document/AllenetalvStateofLouisianaOfficeoftheGovernorDivisionofAdministra?1595341263>
- 2020: In the matter of *Aguilar, Gutierrez, Montes, Palmer and OneAmerica (Plaintiffs) v. Yakima County (Defendant)* in Superior Court of Washington under the recently enacted Washington Voting Rights Act ("WVRA" Wash. Rev. Code § 29A.92.60). In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Providing demographic and analytic litigation support.
  - <https://bloximages.newyork1.vip.townnews.com/yakimaherald.com/content/tncms/assets/v3/editorial/a/4e/a4e86167-95a2-5186-a86c-bb251bf535f1/5f0d01eec8234.pdf.pdf>
- 2018-2020: In the matter of *Flores, Rene Flores, Maria Magdalena Hernandez, Magali Roman, Make the Road New York, and New York Communities for Change (Plaintiffs) v. Town of Islip, Islip Town Board, Suffolk County Board of Elections (Defendants)* in US District Court. On behalf of Defendants - provided a critical analysis of plaintiff's demographic and

environmental justice analysis. The critique revealed numerous flaws in both the demographic analysis as well as the tenets of their environmental justice argument, which were upheld by the court. Ultimately developed mutually agreed upon plan for districting.

- <https://nyelectionsnews.wordpress.com/2018/06/20/islip-faces-section-2-voting-rights-act-challenge/>
- <https://www.courthousenews.com/wp-content/uploads/2018/06/islip-voting.pdf>
- 2017-2020 In the matter of *NAACP, Spring Valley Branch; Julio Clerveaux; Chevon Dos Reis; Eric Goodwin; Jose Vitelio Gregorio; Dorothy Miller; and Hillary Moreau (Plaintiffs) v East Ramapo Central School District (Defendant)* in United States District Court Southern District Of New York (original decision May 25, 2020), later the U.S. Second Circuit Court of Appeals. On behalf of Defendants, developed mutually agreed upon district plan and provided demographic and analytic litigation support.
  - <https://www.lohud.com/story/news/education/2020/05/26/federal-judge-sides-naacp-east-ramapo-voting-rights-case/5259198002/>
  -
- 2017-2020: In the matter of *Pico Neighborhood Association et al (Plaintiffs) v. City of Santa Monica (Defendant)* brought under the California VRA. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Providing demographic and analytic litigation support. Executed geospatial analysis to identify concentrations of Hispanic and Black CVAP to determine the impossibility of creating a minority majority district, and demographic analysis to show the dilution of Hispanic and Black voting strength in a district (vs at-large) system. Work contributed to Defendants prevailing in landmark ruling in the State of California Court of Appeal, Second Appellate District.
  - <https://www.santamonica.gov/press/2020/07/09/santa-monica-s-at-large-election-system-affirmed-in-court-of-appeal-decision>
- 2019: In the matter of *Johnson (Plaintiffs) v. Ardoin / the State of Louisiana (Defendants)* in United States District Court. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Provided demographic and analytic litigation support.
  - <https://www.brennancenter.org/sites/default/files/2019-10/2019-10-16-Johnson%20v%20Ardoin-132-Brief%20in%20Opposition%20to%20MTS.pdf>
- 2019: In the matter of *Suresh Kumar (Plaintiffs) v. Frisco Independent School District et al. (Defendants)* in United States District Court. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Provided demographic and analytic litigation support. Successfully defended.
  - <https://www.friscoisd.org/news/district-headlines/2020/08/04/frisco-isd-wins-voting-rights-lawsuit>
  - <https://www.courthousenews.com/wp-content/uploads/2020/08/texas-schools.pdf>

- 2019: At the request of the City of Frisco, TX in collaboration with demographic testifying expert Dr. Peter Morrison. Provided demographic assessment of the City's potential liability regarding a potential Section 2 Voting Rights challenge.
- 2019: In the matter of *Vaughan (Plaintiffs) v. Lewisville Independent School District et al. (Defendants)* in United States District Court. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Provided demographic and analytic litigation support.
  - <https://www.nbcdfw.com/news/local/lawsuit-filed-against-lewisville-independent-school-district/1125/>
- 2019: In the matter of *Holloway, et al. (Plaintiffs) v. City of Virginia Beach (Defendants)* in United States District Court, Eastern District of Virginia. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Provided demographic and analytic litigation support.
  - <https://campaignlegal.org/cases-actions/holloway-et-al-v-city-virginia-beach>
- 2018: At the request of Kirkland City, Washington in collaboration with demographic testifying expert Dr. Peter Morrison. Performed demographic studies to inform the City's governing board's deliberations on whether to change from at-large to single-member district elections following enactment of the Washington Voting Rights Act. Analyses included gauging the voting strength of the City's Asian voters and forming an illustrative district concentrating Asians; and compared minority population concentration in pre- and post-annexation city territory.
  - [https://www.kirklandwa.gov/Assets/City+Council/Council+Packets/021919/8b\\_SpecialPresentations.pdf#:~:text=RECOMMENDATION%3A%20It%20is%20recommended%20that%20City%20Council%20receive,its%20Councilmembers%20on%20a%20citywide%2C%20at-%20large%20basis.](https://www.kirklandwa.gov/Assets/City+Council/Council+Packets/021919/8b_SpecialPresentations.pdf#:~:text=RECOMMENDATION%3A%20It%20is%20recommended%20that%20City%20Council%20receive,its%20Councilmembers%20on%20a%20citywide%2C%20at-%20large%20basis.)
- 2018: At the request of Tacoma WA Public Schools in collaboration with demographic testifying expert Dr. Peter Morrison. Created draft concept redistricting plans that would optimize minority population concentrations while respecting incumbency. Client will use this plan as a point of departure for negotiating final boundaries among incumbent elected officials.
- 2018: At the request of the City of Mount Vernon, Washington., in collaboration with demographic testifying expert Dr. Peter Morrison. Prepared a numerous draft concept plans that preserves Hispanics' CVAP concentration. Client utilized draft concept redistricting plans to work with elected officials and community to agree upon the boundaries of six other districts to establish a proposed new seven-district single-member district plan.
- 2017: In the matter of *John Hall, Elaine Robinson-Strayhorn, Lindora Toudle, Thomas Jerkins, (Plaintiffs) v. Jones County Board Of Commissioners (Defendant)*. In collaboration with

demographic testifying expert Dr. Peter Morrison. Worked to create draft district concept plans to resolve claims of discrimination against African Americans attributable to the existing at-large voting system.

- <http://jonescountync.gov/vertical/sites/%7B9E2432B0-642B-4C2F-A31B-CDE7082E88E9%7D/uploads/2017-02-13-Jones-County-Complaint.pdf>
- 2017: In the matter of *Harding (Plaintiffs) v. County of Dallas (Defendants)* in U.S. District Court. In collaboration with demographic testifying expert Dr. Peter Morrison. In a novel case alleging discrimination *against* White, non-Hispanics under the VRA, I was retained by plaintiffs to create redistricting scenarios with different balances of White-non-Hispanics, Blacks and Hispanics. Deposed and provided expert testimony on the case.
  - <https://www.courthousenews.com/wp-content/uploads/2018/08/DallasVoters.pdf>
- 2016: Retained by The Equal Voting Rights Institute to evaluate the Dallas County Commissioner existing enacted redistricting plan. In collaboration with demographic testifying expert Dr. Peter Morrison, the focus of our evaluation was twofold: (1) assess the failure of the Enacted Plan (EP) to meet established legal standards and its disregard of traditional redistricting criteria; (2) the possibility of drawing an alternative Remedial Plan (RP) that did meet established legal standards and balance traditional redistricting criteria.
  - <http://equalvotingrights.org/wp-content/uploads/2015/01/Complaint.pdf>
- 2016: In the matter of *Jain (Plaintiffs) v. Coppell ISD et al (Defendant)* in US District Court. In collaboration with demographic testifying expert Dr. Peter Morrison. Consulted in defense of Coppell Independent School District (Dallas County, TX) to resolve claims of discriminatory at-large voting system affecting Asian Americans. While Asians were shown to be sufficiently numerous, I was able to demonstrate that they were not geographically concentrated - thus successfully proving the Gingles 1 precondition could not be met resulting the complaint being withdrawn.
  - <https://dockets.justia.com/docket/texas/txndce/3:2016cv02702/279616>
- 2016: In the matter of *Feldman et al (Plaintiffs) v. Arizona Secretary of State's Office et al, (Defendant)* in SCOTUS. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Provided analytics on the locations and proximal demographics of polling stations that had been closed subsequent to *Shelby County v. Holder* (2013) which eliminated the requirement of state and local governments to obtain federal preclearance before implementing any changes to their voting laws or practices. Subsequently provided expert point of view on disparate impact as a result of H.B. 2023. Advised Maricopa County officials and lead counsel on remediation options for primary polling place closures in preparation for 2016 elections.
  - <https://arizonadailyindependent.com/2016/04/05/doj-wants-information-on-maricopa-county-election-day-disaster/>.

- [https://www.supremecourt.gov/DocketPDF/19/19-1257/142431/20200427105601341\\_Brnovich%20Petition.pdf](https://www.supremecourt.gov/DocketPDF/19/19-1257/142431/20200427105601341_Brnovich%20Petition.pdf)
- 2016: In the matter of *Glatt (Plaintiff) v. City of Pasco, et al. (Defendants)* in US District Court (Washington). In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Provided analytics and draft plans in defense of the City of Pasco. One draft plan was adopted, changing the Pasco electoral system from at-large to a six-district + one at large.
  - <https://www.pasco-wa.gov/DocumentCenter/View/58084/Glatt-v-Pasco---Order---January-27-2017?bidId=>
  - <https://www.pasco-wa.gov/923/City-Council-Election-System>
- 2015: In the matter of *The League of Women Voters et al. (Plaintiffs) v. Ken Detzner et al (Defendants)* in the Florida Supreme Court. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Performed a critical review of Florida state redistricting plan and developed numerous draft concept plans.
  - <http://www.miamiherald.com/news/politics-government/state-politics/article47576450.html>
  - [https://www.floridasupremecourt.org/content/download/322990/2897332/file/OP-SC14-1905\\_LEAGUE%20OF%20WOMEN%20VOTERS\\_JULY09.pdf](https://www.floridasupremecourt.org/content/download/322990/2897332/file/OP-SC14-1905_LEAGUE%20OF%20WOMEN%20VOTERS_JULY09.pdf)
- 2015: In the matter of *Evenwel, et al. (Plaintiffs) v. Abbott / State of Texas (Defendants)* in SCOTUS. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Plaintiffs. Successfully drew map for the State of Texas balancing both total population from the decennial census and citizen population from the ACS (thereby proving that this was possible). We believe this may be the first and still only time this technical accomplishment has been achieved in the nation at a state level. Coauthored SCOTUS Amicus Brief of Demographers.
  - [https://www.supremecourt.gov/opinions/15pdf/14-940\\_ed9g.pdf](https://www.supremecourt.gov/opinions/15pdf/14-940_ed9g.pdf)
  - <https://www.scotusblog.com/wp-content/uploads/2015/08/Demographers-Amicus.pdf>
- 2015: In the matter of *Ramos (Plaintiff) v. Carrollton-Farmers Branch Independent School District* (Defendant) in US District Court (Texas). In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Used 2009-2013 5-year ACS data to generate small-area estimates of minority citizen voting age populations and create a variety of draft concept redistricting plans. Case was settled decision in favor of a novel cumulative voting system.
  - [https://starlocalmedia.com/carrolltonleader/c-fb-isd-approves-settlement-in-voting-rights-lawsuit/article\\_92c256b2-6e51-11e5-adde-a70cbe6f9491.html](https://starlocalmedia.com/carrolltonleader/c-fb-isd-approves-settlement-in-voting-rights-lawsuit/article_92c256b2-6e51-11e5-adde-a70cbe6f9491.html)
- 2015: In the matter of *Glatt (Plaintiff) v. City of Pasco et al. (Defendants)* in US District Court (Washington). In collaboration with demographic testifying expert Dr. Peter Morrison, on



behalf of Defendants. Consulted on forming new redistricting plan for city council review. One draft concept plan was agreed to and adopted.

- <https://www.pasco-wa.gov/923/City-Council-Election-System>
- 2015: At the request of Waterbury, Connecticut, in collaboration with demographic testifying expert Dr. Peter Morrison. As a result of a successful ballot measure to convert Waterbury from an at-large to a 5-district representative system -consulted an extensive public outreach and drafted numerous concept plans. The Waterbury Public Commission considered alternatives and recommended one of our plans, which the City adopted.
  - <http://www.waterburyobserver.org/wod7/node/4124>
- 2014-15: In the matter of *Montes (Plaintiffs) v. City of Yakima (Defendant)* in US District Court (Washington). In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants. Analytics later used to support the Amicus Brief of the City of Yakima, Washington in the U.S. Supreme Court in *Evenwel v. Abbott*.
  - <https://casetext.com/case/montes-v-city-of-yakima-3>
- 2014: At the request of Gulf County, Florida in collaboration with demographic testifying expert Dr. Peter Morrison. Upon the decision of the Florida Attorney General to force inclusion of prisoners in redistricting plans – drafted numerous concept plans for the Gulf County Board of County Commissioners, one of which was adopted.
  - <http://myfloridalegal.com/ago.nsf/Opinions/B640990E9817C5AB85256A9C00631387>
- 2012-2015: In the matter of *GALEO (Plaintiffs) and the City of Gainesville (Defendants)* in Georgia. In collaboration with demographic testifying expert Dr. Peter Morrison, on behalf of Defendants -consulted on defense of existing at-large city council election system.
  - <http://atlantaprogressivenews.com/2015/06/06/galeo-challenges-at-large-voting-in-city-of-gainesville/>
- 2012-: Confidential. Consulted (through Morrison & Associates) to support plan evaluation, litigation, and outreach to city and elected officials (1990s - mid-2000s). Executed first statistical analysis of the American Community Survey to determine probabilities of minority-majority populations in split statistical/administrative units of geography, as well as the cumulative probabilities of a “false-negative” minority-majority reading among multiple districts.
- 2011-: Confidential. Consulted on behalf of plaintiffs in *Committee (Private) vs. State Board of Elections* pertaining to citizen voting-age population. Evaluated testimony of defense expert, which included a statistical evaluation of Hispanic estimates based on American Community Survey (ACS) estimates. Analysis discredited the defendant’s expert’s analysis and interpretation of the ACS.

### Applied Demography Studies

In addition to political redistricting cases, BGD has provided demographic and analytic expertise across a broad array of issues, oftentimes creating pivotal evidence that has been decisive in legal cases and analytics that were core to the success of clients. Examples include:

- 2018-2019: Client Confidential. Leveraged the National Survey of Drug Use and Health (NSDUH) to develop a comprehensive analysis of opioid use, misuse, dependence, abuse and opioid use disorder (OUD). Analytics included prevalence analysis, demographic profiles, insurance and treatment trends, comorbidities, other drug use, dependence and abuse covariates. This analysis culminated in what is believed to be the only long-term forecast of opioid misuse, dependence, abuse and OUD by age, race and sex, marital status, educational attainment and income for the United States.
- 2016-ongoing: Consulted (through Morrison & Associates) in defense against a US Department of Justice housing discrimination complaint against Oyster Bay, New York. Leveraged 9 years of the ACS PUMS data measure local demographic makeup and gross migration patterns by age and race. Findings refuted plaintiffs' claim that housing practices were discriminatory. (Access at: <http://oysterbayguardian.com/stories/Town-accused-of-housing-discrimination,145>)
- 2016: Consulted (through Morrison & Associates) in defense class action claim against Shorter University (Georgia) in Bishop, et al. v. Shorter University, Inc. Estimated the citizenship/state of residence of the proposed class members using 2008-2014 ACS PUMS data. Findings contributed to partial dismissal of the case. (access at: <http://www.lexislegalnews.com/articles/10983/university-awarded-attorney-fees-for-discovery-disputes-in-stolen-records-suit> )
- 2016: Consulted (through Morrison & Associates) in defense of class action claim against consumer product manufacturer. Used ACS PUMS data to compute gross annual out-migration flows from Illinois. Data were used to calibrate demographic accounting model that tracks movement over time of victim cohorts for whom legal redress is sought.
- 2013 - Consulted (through Morrison & Associates) in using the 2008-2012 ACS and Census 2000 PUMS data to estimate Persons >/< 18 in household by number of bedrooms in a technical analysis supporting housing discrimination litigation sought in Florida.
- 2012 - Consulted (through Morrison & Associates) in generating a time-series of ACS Citizen Voting Age Population estimates by race and ethnicity from 2005-2010 ACS to assess the impact of a State of Wisconsin proposed rule requiring driver licenses to verify voters' current addresses.
- 2011 - Commissioned by ESRI as member of expert team to conduct 1<sup>st</sup>-ever evaluation of small-area private-party population estimates to produce "Vendor Accuracy Study Population and Household Estimates vs. 2010 Census". Built databases containing five vendors across

>200,000 units of geography and executed 4 summary error calculations for each to provide to succeeded experts for analysis and interpretation (with David A. Swanson, Jeff Tayman and Jerome McKibben).

- 2008 - Commissioned as expert to provide demographic analytic support in defense of putative class-action lawsuit against DuPont alleging PFOA's and other chemicals were released by Teflon products during cooking. Used Consumer Expenditure Survey, Current Population Survey, and U.S. Census PUMS data to produce model showing how interstate migration diluted the class over the 10-year class period plaintiffs sought (with Morrison & Associates). Result: class successfully de-certified (access summary at: <http://www.masstortdefense.com/2008/12/articles/federal-court-denies-class-certification-in-teflon-litigation/>).
- 2008 – Retained by RAND Corporation as expert to create a Census 2000 Block-Group (BG) level file that re-allocates the “some other race” population to known race-alone and Hispanic categories. File was generated using U.S. Census Bureau’s county-level modified age-race-sex (MARS) file, U.S. Census 2000 data and the U.S. Census Bureau’s Iterative Proportionate Fitting (IPF) algorithm (with Morrison & Associates).
- 2008 – Commissioned as expert to analyze 2007 US NAICS Manufacturing Codes and Total Employment for Louisiana, Mississippi and Texas in support of local-area forecasting for area school districts (with Cropper G.I.S.).
- 2007 - Consulted (through Morrison & Associates) to provide demographic analytic support of the “Proposed Revision of The Census Bureau’s 2006 Population Estimate for the Town of Nantucket, MA”. Result: successful challenge accepted by the U.S. Census Bureau.
- 2007 - Produced annual birth and death data from U.S. vital statistics and state-specific migration rates from IRS data for 2001-2003 in support of presentation to Western Planners Association (with McKibben Demographics).
- 2007 - Produced California age-specific time series of domestic and international migration from the U.S. Census Modified Age race Sex (MARS) and Census 2000 data files for the State of California (with McKibben Demographics).
- 2007 - Consulted (through Morrison & Associates) to write “Analysis of Census 2000 Hispanic Estimate in Westchester, NY”.
- 2006 - Commissioned as expert to provide demographic analysis of the Island of Hawaii for Tradewinds Forest Products. Hilo, Hawaii.
- 2006 – Commissioned as expert by wine retailer “Vino 100” to support site location in Wisconsin. Used U.S. Consumer Expenditure Survey and U.S. Census data to produce a geospatial model predicting optimal site location in Milwaukee, WI. The recommended site was chosen, and grew to the #2 highest grossing store in the chain in < 12 months.

- 2006 - Consulted (through Morrison & Associates) to use the U.S. Census PUMS data to estimate the proportion of Hispanic and non-Hispanic individuals aged 18+, who are citizens, who speak English well by migration status in Yolo, California.
- 2004 - Consulted (through Morrison & Associates) to provide demographic analytic support of Anheuser Busch in defense of putative class-action lawsuit alleging: “minors’ intentional violations of state alcohol laws on lawful product advertising, generally asserting theories of consumer fraud, unjust enrichment and public nuisance”. Leveraged IRS State-to-State migration flow files and 2000 Census PUMS data file to prove inter-state class dilution (with Morrison & Associates). Result: case dismissed in 2006. (access at: [http://www.wikinvest.com/stock/Anheuser-Busch Companies \(BUD\)/Legal Proceedings](http://www.wikinvest.com/stock/Anheuser-Busch%20Companies%20(BUD)/Legal%20Proceedings))
- 2004 - Consulted (through Morrison & Associates) to provide demographic analytic support of Modesto, CA (defendants) against plaintiff claim of racially-biased annexation practices.
- 2004 - Consulted (through Morrison & Associates) to provide demographic analysis of market potential of “Experience Pennsylvania” tourism campaign.
- 2004 - Commissioned as expert to provide opinion on need and site location for cardiac facility for the Future Forth Valley Healthcare Strategy Initiative. Leveraged Nationwide Inpatient Sample - part of the Healthcare Cost and Utilization Project (HCUP), State of South Carolina inpatient data and U.S. Census data to provide recommendation that was approved (with Third Wave Research).
- 2003 - Commissioned as expert to provide demographic analysis of Antelope Family YMCA “Lancaster Site” location (with Morrison & Associates).
- 2003 – “Real World Business Demography” seminar taught at request of Dr. Roger Hammer RSOC 676 “Applied Demography” at University of Wisconsin.
- 2002 – Commissioned to develop and produce data and methodology using 5-year migration data in support of aging-in-place analytics in Pittsburgh, PA (with Morrison & Associates).

Note: numerous other projects redacted at client’s request or due to confidentiality.

**Note: The remainder of this page is intentionally blank**

School Redistricting and Municipal Infrastructure Projects

BGD worked with McKibben Demographics from 2004-2012 providing expert demographic and analytic support. These engagements involved developing demographic profiles of small areas to assist in building fertility, mortality and migration models used to support long-range population forecasts and infrastructure analysis in the following communities:

Fargo, ND 10/2012	Charleston, SC 8/08
Columbia, SC 3/2012	Woodland, IL 7/08
Madison, MS 9/2011	White County, IN 6/08
Rockwood, MO 3/2011	Gurnee District 56, IL 5/08
Carthage, NY 3/2011	Central Noble, IN 4/08
NW Allen, IN 9/2010	Charleston First Baptist, SC 4/08
Fayetteville, AR 7/2010	Edmond, OK 4/08
Atlanta, GA 2/2010	East Noble, IN 3/08
Caston School Corp., IN 12/09	Mill Creek, IN 5/06
Rochester, IN 12/09	Rhode Island 5/06
Urbana, IL 11/09	Garrett, IN 3/08
Dekalb, IL 11/09	Meridian, MS 3/08
Union County, NC 11/09	Madison County, MS 3/08
South Bend, IN 8/09	Charleston 12/07
Lafayette, LA 8/09	Champaign, IL 11/07
Fayetteville, AR 4/09	Richland County, SC 11/07
New Orleans, LA 4/09	Lake Central, IN 11/07
Wilmington New Hanover 3/09	Columbia, SC 11/07
New Berry, SC 12/08	Duneland, IN 10/07
Corning, NY 11/08	Union County, NC 9/07
McLean, IL 11/08	Griffith, IN 9/07
Lakota 11/08	Rensselaer, IN 7/07
Greensboro, NC 11/08	Hobart, IN 7/07
Guilford 9/08	Buffalo, NY 7/07
Lexington, SC 9/08	Oak Ridge, TN 5/07
Plymouth, IN 9/08	Westerville, OH 4/07

Projects Continued

Baton Rouge, LA 4/07	Allen County 11/05
Cobb County, GA 4/07	Bremen, IN 11/05
Charleston, SC District 20 4/07	Smith Green, IN 11/05
McDowell County, NC 4/07	Steuben, IN 11/05
East Allen, IN 3/07	Plymouth, IN 11/05
Mt. Pleasant, SC District 2 2/07	North Charleston, SC 11/05
Peach County, GA 2/07	Huntsville, AL 10/05
North Charleston, SC District 4 2/07	Dekalb, IN 9/05
Madison County, MS revisions 1/07	East Noble, IN 9/05
Portage County, IN 1/07	Valparaiso, IN 6/05
Marietta, GA 1/07	Penn-Harris-Madison, IN 7/05
Porter, IN 12/06	Elmira, NY 7/05
Harrison County, MS 9/06	South Porter/Merrville, IN 7/05
New Albany/Floyd County, IN 9/06	Fargo, ND 6/05
North Charleston, SC 9/06	Washington, IL 5/05
Fairfax, VA 9/06	Addison, NY 5/05
Coleman 8/06	Kershaw, SC 5/05
DeKalb, GA 8/06	Porter Township, IN 3/05
LaPorte, IN 7/06	Portage, WI 1/05
NW Allen, IN 7/06	East Stroudsburg, PA 12/04
Brunswick, NC 7/06	North Hendricks, IN 12/04
Carmel Clay, IN 7/06	Sampson/Clinton, NC 11/04
Calhoun, SC 5/06	Carmel Clay Township, IN 9/04
Hamilton Community Schools, IN 4/06	SW Allen County, IN 9/04
Dilworth, MN 4/06	East Porter, IN 9/04
Hamilton, OH 2/06	Allen County, IN 9/04
West Noble, IN 2/06	Duplin, NC 9/04
New Orleans, LA 2/06	Hamilton County / Clay TSP, IN 9/04
Norwell, IN 2/06	Hamilton County / Fall Creek TSP, IN 9/04
Middletown, OH 12/05	Decatur, IN 9/04
West Noble, IN 11/05	Chatham County / Savannah, GA 8/04
Madison, MS 11/05	Evansville, IN 7/04
Fremont, IN 11/05	Madison, MS 7/04
Concord, IN 11/05	Vanderburgh, IN 7/04
	New Albany, IN 6/04

### **Publications**

- “Constructing Life Tables from the Kaiser Permanente Smoking Study and Applying the Results to the Population of the United States” Population Research & Policy Review (with Dr. Dave Swanson and Dr. Simeon Chow) 2020.
- Peter A. Morrison and Thomas M. Bryan, Redistricting: A Manual for Analysts, Practitioners, and Citizens (2019). Springer Press: Cham Switzerland.
- “Small Area Business Demography.” in D. Poston (editor) Handbook of Population, 2<sup>nd</sup> Edition. (2018). Springer Press: London (with D. Swanson and S. Smith).
- “From Legal Theory to Practical Application: A How-To for Performing Vote Dilution Analyses.” Social Science Quarterly. (with M.V. Hood III and Peter Morrison). March 2017 <http://onlinelibrary.wiley.com/doi/10.1111/ssqu.12405/abstract>
- In the Supreme Court of the United States Sue Evenwel, Et Al., *Appellants*, V. Greg Abbott, in his official capacity as Governor of Texas, et al., *Appellees*. *On appeal from the United States District Court for the Western District of Texas*. *Amicus Brief of Demographers Peter A. Morrison, Thomas M. Bryan, William A. V. Clark, Jacob S. Siegel, David A. Swanson, and The Pacific Research Institute - As amici curiae* in support of Appellants. August 2015. (access at: [www.scotusblog.com/wp-content/uploads/2015/08/Demographers-Amicus.pdf](http://www.scotusblog.com/wp-content/uploads/2015/08/Demographers-Amicus.pdf))
- Workshop on the Benefits (and Burdens) of the American Community Survey, Case Studies/Agenda Book 6 “Gauging Hispanics’ Effective Voting Strength in Proposed Redistricting Plans: Lessons Learned Using ACS Data.” June 14–15, 2012 <http://docplayer.net/8501224-Case-studies-and-user-profiles.html>
- “MAPE-R: A Rescaled Measure of Accuracy for Cross-Sectional, Sub-national Forecasts.” Journal of Population Research 28: 225-243 (with Dr. Dave Swanson and Dr. Jeff Tayman). 2011.
- “Targeting Spatial Clusters of Elderly Consumers in the U.S.” in Population Research & Policy Review. Access at: <http://link.springer.com/article/10.1007/s11113-009-9149-2>
- “Basic Sources of Statistics” by Bryan, Thomas in J. Siegel and D. Swanson (eds.) The Methods and Materials of Demography, Condensed Edition, Revised. (2004). Academic/Elsevier Press: Los Angeles (with D. Swanson and P. Morrison).
- “Collection and Processing of Demographic Data” by Bryan, Thomas in J. Siegel and D. Swanson (eds.) The Methods and Materials of Demography, Condensed Edition, Revised. (2004). Academic/Elsevier Press: Los Angeles (with D. Swanson and P. Morrison).
- “Internal and Short Distance Migration” by Bryan, Thomas in J. Siegel and D. Swanson (eds.) The Methods and Materials of Demography, Condensed Edition, Revised. (2004). Academic/Elsevier Press: Los Angeles (with D. Swanson and P. Morrison).

- "Population Estimates" by Bryan, Thomas in J. Siegel and D. Swanson (eds.) The Methods and Materials of Demography, Condensed Edition, Revised. (2004). Academic/Elsevier Press: Los Angeles (with D. Swanson and P. Morrison).
- "Demographic Trends and Market Analysis: Knowledge is Power, Market Analysis 101." *Shopping Center Business.* May, 2002.
- "U.S. Census Bureau Population Estimates and Evaluation with Loss Functions." *Statistics in Transition Journal.* Warsaw, Poland. March 2000.
- Yucca Mountain Site Characterization Project: Summary of Socioeconomic Data Analyses Conducted in Support of the Radiological Monitoring Program: April 1997 to April 1998. TRW Environmental Safety Systems, Inc. Las Vegas, Nevada (with Dave Swanson). June 1998.
- Yucca Mountain Site Characterization Project: Summary of Socioeconomic Data Analyses Conducted in Support of the Radiological Monitoring Program: April 1996 to April 1997. TRW Environmental Safety Systems, Inc., Las Vegas, Nevada (with Dave Swanson). June 1997.
- "The Size of Selected Lifestyle Segments: 1990 to 2010." Third Wave Research, Madison, WI. (with D. Swanson and G. Hough) March 1996.
- "Population Estimation Techniques Using the Housing Unit Method." Master of Urban Science (M.U.S.) Research Paper. Department of Urban Studies, Portland State University (Co-chaired by D. Swanson and George Hough). June 1996.

#### **Professional Presentations and Conference Participation**

- "New Technical Challenges in Post-2020 Redistricting" 2020 Population Association of America Applied Demography Conference, 2020 Census Related Issues, February 2021. With Dr. Peter Morrison. <https://www.youtube.com/watch?v=ETvvoECt9sc&feature=youtu.be>
- "Tutorial on Local Redistricting" 2020 Population Association of America Applied Demography Conference, February 2021. With Dr. Peter Morrison. <https://www.youtube.com/watch?v=ETvvoECt9sc&feature=youtu.be>
- "Demographic Constraints on Minority Voting Strength in Local Redistricting Contexts" 2019 Southern Demographic Association meetings (coauthored with P. Morrison) New Orleans, LA, October 2019.
- "The Implications of Demography Trends for Future Opioid Abuse," 2019 Southern Demographic Association meetings (coauthored with Dr. Rick Thomas) New Orleans, LA, October 2019.
- "Prisoner Populations and Redistricting: Counting vs. Discounting," 2019 Southern Demographic Association meetings (coauthored with P. Morrison) New Orleans, LA, October 2019.



- "Estimating the Potential Population Health Impact of Authorizing the Marketing of E-cigarettes in the US" with Muhammad-Kah, R.; Hannel, T.; Wei, L.; Black, R.; Gogova, M.; Pithawalla, Y.B.; Presented at 24th Annual Meeting of the Society for Research on Nicotine and Tobacco (SRNT), Baltimore, Maryland, February 21-24, 2018.
- "Estimating the Population Health Impact of Authorizing the Marketing of a Smokeless Tobacco Product with a Proposed Modified Risk Claim" with Muhammad-Kah, R.; Hannel, T.; Wei, L.; Black, R.; Gogova, M.; Pithawalla, Y.B.; 24th Annual Meeting of the Society for Research on Nicotine and Tobacco (SRNT), Baltimore, MD, February 21-24, 2018.
- "The Impact of Tobacco Use History on e-Cigarette and Cigarette Transition Patterns - A Longitudinal Analysis of Population Assessment of Tobacco and Health (PATH) Study" with Wei, L.; Black, R.; Muhammad-Kah, R.; Pithawalla, Y.B.; Chow, S.; 24th Annual Meeting of the Society for Research on Nicotine and Tobacco (SRNT), Baltimore, MD, February 21-24, 2018.
- "Projecting Future Demand for Assisted Living: A Case Study" 2017 Population and Public Policy Conference, Houston, TX.
- "Applications of Big Demographic Data in Running Local Elections" 2017 Population and Public Policy Conference, Houston, TX.
- "Distinguishing 'False Positives' Among Majority-Minority Election Districts in Statewide Congressional Redistricting," 2017 Southern Demographic Association meetings (coauthored with P. Morrison) Morgantown, WV.
- "Devising a Demographic Accounting Model for Class Action Litigation: An Instructional Case" 2016 Southern Demographic Association (with Peter Morrison), Athens, GA.
- "Gauging Hispanics' Effective Voting Strength in Proposed Redistricting Plans: Lessons Learned Using ACS Data." 2012 Conference of the Southern Demographic Association, Williamsburg, VA.
- "MAPE-R: An Empirical Assessment." 2011 Conference of the Population Association of American (with Jeff Tayman and Dave Swanson) Washington, D.C.
- "MAPE-R: A Refined Measure of Accuracy for Ex Post Evaluation of Estimates and Forecasts." Presented at the 2010 International Symposium of Forecasting (with J. Tayman and D. Swanson) San Diego, CA.
- "Targeting Spatial Clusters of Elderly Consumers in the U.S." Co-authored for the 2007 International Seminar on Applications of Demography in Business (presented by Peter Morrison) Sydney, Australia.
- "Characteristics of the Arab-American Population from Census 2000 and 1990: Detailed Findings from PUMS." 2004 Conference of the Southern Demographic Association, (with Samia El-Badry) Hilton Head, SC.

- "Small-Area Identification of Arab American Populations," 2004 Conference of the Southern Demographic Association, Hilton Head, SC.
- "New Approaches to Spotting Elderly Enclaves." 2004 Conference of the Population Association of America, (with Peter Morrison) Boston, MA.
- "Small Area Market Potential of Hospitals." 2004 Conference of the Population Association of America, Boston, MA.
- "Spatial Research Frontiers Using GIS" 2002 Southern Demographic Association, Austin, TX.
- "MAPE-R: It's Features and Results from a National Block-Group Test." 2002 Conference of the American Statistical Association. (with D. Swanson, J. Tayman, and C. Barr). New York City, NY
- "Applied Demography in Action: A Case Study of Population Identification." 2002 Conference of the Population Association of America, Atlanta, GA.
- "CACI One" Product Presentation/Poster Session, presented at the 2000 Conference of the Population Association of America, Washington, DC.
- "Statistical Evaluation of Distributive Housing Unit Method" 2000 Conference of the Southern Demographic Association, New Orleans, LA.
- "Results of FSCPE Survey on Small-Area Estimate Accuracy and the Development of Data Mining Techniques to Detect Problematic Cases in Small Area Estimates" 2000 Conference of the Population Association of America, Los Angeles, CA.
- "Small Area Population Estimates Methodology in the United States." 1999 Conference of the Population Association of America, New York, NY.
- "On the Measurement of Accuracy for Subnational Demographic Estimates Using MAPE Transformation and Re-Expressions." Presented at the 1999 U.S. Census Bureau Population Estimates Methods Conference (with D. Swanson, J. Tayman and C. Barr) Washington, D.C.
- "U.S. Census Bureau Estimates and Evaluation with Loss Functions" 1999 Conference of the International Statistics Institute, Helsinki, Finland.
- "Evaluating Estimate Outliers with Loss Functions." 1999 Conference of the International Association of Survey Statisticians, Riga, Latvia.
- "Evaluation of 1998 Subcounty Population Estimates." 1999 Conference of the Federal-State Cooperative for Population Estimates, Baltimore, MD.
- "Evaluation of Components of the Housing Unit Method." 1999 Conference of the Southern Demographic Association," San Antonio, TX.

- "Small-Area Population Estimation Technique Using Administrative Records and Evaluation of Results with Loss Functions and Optimization Criteria." 1999 Conference of the Federal Committee on Statistical Methodology, Arlington, VA.
- "Housing Unit Estimates and Estimates Geography." 1998 Conference of the Federal-State Cooperative for Population Estimates, Park City, UT.
- "A Test of the Housing Unit Method in Multnomah County, OR." 1997 Conference of the Population Association of America, Washington, DC.
- "Linear and Logarithmic Population Forecasting Techniques." 1996 Conference of the Federal-State Cooperative for Population Projections, New Orleans, LA.
- "Small Area Population Estimates Using the Housing Unit Method." 1996 Conference of the Southern Demographic Association, Memphis, TN.

**Professional Conference Chairs, Peer Reviews and Conference Discussant Roles**

- "The Historical Roots of Contentious Litigation Over Census Counts in the Late 20th Century". Peer reviewer for presentation at the Hawaii International Conference on the Social Sciences, Honolulu, Hawaii, June 17-19, 2004 by David A. Swanson and Paula A. Walashek.
- 2004 - Population Research and Policy Review External Peer Reviewer / MS #253 "A New Method in Local Migration and Population Estimation".
- Session Discussant on "Spatial Demography" at the 2003 Conference of the Southern Demographic Association, Arlington, VA.
- Subject Moderator at the International Program Center (IPC) 2000 Summer Workshop on Subnational Population Projections for Planning, Suitland, MD.
- Session Chairman on "Population Estimates: New Evaluation Studies" at the Conference of the Southern Demographic Association, Austin, TX.
- Conference Session Chairman at the 2000 Conference of the Federal Forecasters Conference (FFC), Washington, DC.
- Session Discussant on "New Developments in Demographic Methods" at the 2000 Conference of the Southern Demographic Association, New Orleans, LA.
- Panel Discussant on GIS Applications in Population Estimates Review at the 2000 Conference of the Population Association of America, Los Angeles, CA.
- Panel Discussant on Careers in Applied Demography at the 2000 Conference of the Population Association of America, Los Angeles, CA.

**Professional Employment History (now retired)**

**June 2019-May 2020: Swedish Match North America / Senior Director: Marketing Research and Analytics**

Responsibilities: reported to SMNA executive leadership and directed the development and execution of adult consumer research and enhancement of the business intelligence function. My objectives were to build an Analytic and Research Center of Excellence, to develop analytic and leadership talent within the organization and to create rigorous and repeatable processes and drive the development and success of reduced-harm alternatives to cigarettes. Led the development and execution of market and consumer research and reporting for ZYN and other smokeless / reduced harm tobacco products.

**December 2012-February 2019: Altria Center of Excellence / Director: Population Modeling, Consumer Tracking and Analytics**

Responsibilities: directed the adult consumer research and advanced analytic function for the Altria OpCos (PMUSA, USSTC and NuMark), Regulatory Affairs and the Food and Drug Administration (RA/FDA) engagements.

*OpCo* engagement included managing adult consumer tracking infrastructure to deliver timely and accurate reads of adult consumer behavior in the marketplace; and marketing science and survey research design and advanced analytic support across Altria's marketing function.

*RA/FDA* engagement included acquiring and managing health data, the development and execution of population modeling infrastructure to evaluate the health impacts of introducing reduced-harm tobacco products, and the development of postmarket surveillance tools to measure the performance of reduced-harm tobacco products in the marketplace.

- Internal clients: Altria's Executive Leadership Team; Legal (provided litigation support); HR (drawing upon my diversity and inclusion expertise); Business Analysis and Research; Investor Relations and External Affairs; Brand, Strategy & Business Development (to support Merger & Acquisition activity); and Forecasting & Business Analysis (drawing upon my advanced analytic expertise).
- Management responsibilities: five-member staff of senior analysts / managers and two offshore KPO analytic and research teams (11 FTE total).
- Connecting Altria with external expertise as needed, based upon my peer network of academic researchers.

**May 2011-November 2012: Microsoft / Senior Manager: Central Marketing Group MS Office**

Responsibilities: managed the global market research of consumers and small- to mid-market businesses for Microsoft Office 365 release. Notable accomplishments:

- **O365 Feature and Functionality Testing:** Managed all aspects (including budgeting, contract negotiation, research execution and reporting) of online research and subsequent Kano analysis among O365 target audiences. Research results were used to identify which O365 features and functionality would be used throughout the subsequent value proposition work and featured in the O365 media campaign.
- **Office Impact Tracker (OIT):** Managed all aspects of the biannual Office tracker covering four key objectives: 1) assessing the current state of the Microsoft Office business; 2) measuring multiple device ownership; 3) interpreting productivity tasks and usage scenarios on devices; and 4) gauging [measuring] the size of the consumer and small business markets. Research was successfully executed on time and under budget in the US, France, Germany and Brazil.
- **SMB Segmentation:** Managed the design and execution of a latent class segmentation to identify the major firmographic, attitudinal and behavioral differences across SMB. Managed additional qualitative research to develop personas for each segment. Resulting model now serves as the foundation for other current and future SMB research at Microsoft and is currently used as the targeting vehicle for upcoming Microsoft campaigns. This research was executed on time and under budget in the US, Germany, Korea, India and Brazil.

May 2005-May 2011: Altria / Manager of Market Information and Consumer Research

Responsibilities: developed and enhanced advanced consumer research and business analytics. Oversee Information Management and Forecasting / Business Analysis groups. Notable accomplishments:

- Managed one of the nation's largest and most complex (multi-audience / multi-mode) adult CPG tracking surveys – including contracting a multi-year, multi-million dollar agreement for continued service delivery.
- Developed marketing science models to leverage product concept purchase interest scores into post-launch share-of-category forecasts.
- Authored Altria's organizational consumer research supplier management handbook.
- Developed advanced analytics and numerous predictive models in support of the OpCos.

April 2003-May 2005 Third Wave Research / Director: Population Research

Responsibilities: managed corporate G.I.S., developed demographic and business data, managed customer accounts and implemented statistical software development standards. Performed B-2-B and B-2-C customer analyses that integrated primary survey and research data with consumer household and business databases, US Census data, and other secondary data sources. Experience with large databases, sampling and processing of survey research data. Notable accomplishments:

- Built a nationwide census block-group level population estimate & forecast system, used for integrated targeted marketing.

- Built improved methods for estimating the small-area market potential of hospitals using the Nationwide Inpatient Sample (part of the Healthcare Cost and Utilization Project / HCUP).
- Built the first National Basketball Association season ticket holder customer segmentation.

January 2001- April 2003 ESRI Business Information Solutions / Demographer

Responsibilities included demographic data management, small-area population forecasting, IS management and software product and specification development. Additional responsibilities included developing GIS-based models of business and population forecasting, and analysis of emerging technology and R&D / testing of new GIS and geostatistical software.

May 1998-January 2001 U.S. Census Bureau / Statistician

Responsibilities: developed and refined small area population and housing unit estimates and innovative statistical error measurement techniques, such as Loss Functions and MAPE-R.

**Primary Software Competencies**

ESRI ArcGIS: advanced

SAS: intermediate

Microsoft Office: advanced

**Professional Affiliations**

International Association of Applied Demographers (IAAD) Board of Directors

Population Association of America (Member)

Southern Demographic Association (Member)

American BAR Association (Affiliated Professional: Solo, Small Firm and General Practice Division)

**Service**

Eagle Scout, 1988, Boy Scouts of America. Member of the National Eagle Scout Association. Involved in leadership of the Boy Scouts of America Heart of Virginia Council.

Prior Director, "Salute" Recruitment and Development – Altria's external engagement group with US Veterans.



**References**

Dr. David Swanson  
*Professional Peer*  
[david.swanson@ucr.edu](mailto:david.swanson@ucr.edu)  
951-534-6336

Dr. Jerome McKibben  
*Professional Peer*  
[j.mckibben@mckibbendemographics.com](mailto:j.mckibben@mckibbendemographics.com)  
978-501-7069

Dr. Peter Morrison  
*Professional Peer*  
[petermorrison@me.com](mailto:petermorrison@me.com)  
310-266-9580

Brian Cruikshank  
*President, Engine Insights North America*  
[brian.cruikshank@enginegroup.com](mailto:brian.cruikshank@enginegroup.com)  
612-205-4846



**Mohamadi Sarkar**  
Director of Regulatory Sciences at Altria  
August 9, 2017, Mohamadi was senior to Thomas but didn't manage directly

Tom is one of the smartest numbers guys that I have met. He has the unique ability of dissecting complex data analysis into easily understandable and actionable outcomes. He is a true team player and great colleague to work with.



**Peter A. Morrison**  
Applied Demographic Analysis  
August 8, 2017, Thomas worked with Peter A. in the same group

Tom is a superb data scientist, with deep experience accessing and using Census and other public data. (He was once a statistician at the Census Bureau.) Tom adheres to standards distinctive of any seasoned Census Bureau statistician, notably understanding how to assure quality control with "big data." Most of our projects here at Morrison & Associates rely heavily on Tom's analytic skills, GIS proficiency, and overall years of experience.



**Ozlem Yaylaci** • 1st  
Associate Manager & Applied Statistician at Altria

Thank you for your leadership & support Tom. You are the best manager I've ever had and I learned a lot from you. Good luck!



Tom, thank you sincerely for your leadership, constant support, wisdom, and reliable direction during this busy and uncertain time. You have done an incredible job shaping our team into a fully integrated and supportive unit.

Alex Ogilvie

*"Flectere si nequeo superos, Acheronta movebo"*

# **EXHIBIT B**



## Second Expert Report of Michael Barber

Dr. Michael Barber  
Brigham Young University  
724 Spencer W. Kimball Tower  
Provo, UT 84604  
barber@byu.edu

25 March 2021

# 1 Introduction and Qualifications

I am an associate professor of political science at Brigham Young University and faculty fellow at the Center for the Study of Elections and Democracy in Provo, Utah. I received my PhD in political science from Princeton University in 2014 with emphases in American politics and quantitative methods/statistical analyses. My dissertation was awarded the 2014 Carl Albert Award for best dissertation in the area of American Politics by the American Political Science Association.

I teach a number of undergraduate courses in American politics and quantitative research methods.<sup>1</sup> These include classes about political representation, Congressional elections, statistical methods, and research design.

I have worked as an expert witness in a number of cases in which I have been asked to perform and evaluate various statistical methods. Cases in which I have testified at trial or by deposition are listed in my CV, which is attached to the end of my initial report, dated March 9, 2021.

In my position as a professor of political science, I have conducted research on a variety of election- and voting-related topics in American politics and public opinion. Much of my research uses advanced statistical methods for the analysis of quantitative data. I have worked on a number of research projects that use “big data” that include millions of observations, including a number of state voter files, campaign contribution lists, and data from the US Census.

Much of this research has been published in peer-reviewed journals. I have published nearly 20 peer-reviewed articles, including in our discipline’s flagship journal, *The American Political Science Review* as well as the inter-disciplinary journal, *Science Advances*. My CV details my complete publication record.

The analysis and explanation I provide in this report are consistent with my training in statistical analysis and are well-suited for this type of analysis in political science and

---

<sup>1</sup>The political science department at Brigham Young University does not offer any graduate degrees.

quantitative analysis more generally.

I have been asked to evaluate and explain at an approachable level the process of differential privacy (DP), its application to the 2020 Census, and how it fits within the field of probability theory and statistical methods.

## 2 The Process of Differential Privacy in the US Census

This section provides a very basic explanation of the differential privacy and post-processing procedure that the US Census Bureau plans to implement in the 2020 Census and how the process is a straightforward application of common statistical methods. The process can be divided into three basic steps. While the application of these steps across millions of geographic units and sub-groups of the population requires complicated mathematical and statistical methods as well as immense computational capacity, the concepts are in fact quite simple to describe.<sup>2</sup>

The Census Bureau argues that their method of differential privacy and post-processing does not fall inside the definition of statistical inference because they are not using “the drawing of inferences about a population based on data taken from a sample of that population (pg. 7 of Department of Commerce reply).” However, this definition of statistical inference is overly narrow. Statistical inference also refers to other processes aside from the definition provided by the Department of Commerce. Researchers often use datasets that include the entire population of data and still make inferences, or comparisons that are intended to inform us of differences that exist across the population. For example, suppose I had health information for the entire United States population and was looking at the variation in rates of heart disease. From these data I might learn that there are large differences across the country geographically in the rate of heart disease, as well as differences based on various demographic traits. Furthermore, I might then draw comparisons between the geographic

---

<sup>2</sup>This description is not intended to be a complete nor technical explanation of the differential privacy and post-processing procedure. Nevertheless, the basic principles outlined here are helpful in understanding how the process works.

differences versus the demographic variation. These inferences are no less “statistical inference” because they came from the population rather than a sample of the population. While our discussion of the variation would differ from a discussion of statistical uncertainty that comes with using samples of data (and the associated sampling error), we would nonetheless still be interested in the variation associated with race, or age, or some other trait compared to the natural variation that occurs across other features of the population. Thus, statistical inference can also include making comparisons across a population, and not just a sample.

This applies to the differential privacy and post-processing method proposed by the Census Bureau because they are engaged in a similar process as described above. Using data on the entire population, they are using a sophisticated statistical algorithm to learn about differences, or variation, in the population. In this case, they are interested in variation in demographic parameters across the country that might lead to leakages of privacy. Once those groups, or subgroups, of people have been identified, they then apply the parameters of their model to inject noise and further adjust that noise via post-processing to produce the confidential dataset. Evens et. al (2020) describe the process in the following way: “privacy researchers typically begin with the choice of a target (confidential) *dataset*, add *privacy-protective procedures*, and then use the resulting *differentially private dataset or analyses* to infer to the confidential dataset (pg. 3, emphasis in original).”<sup>3</sup> Thus the process of differential privacy and post-processing is using information from the population that inform the choice of probability distributions that are then sampled from to generate noise that creates a confidential dataset that infers, or is a “noisy” estimate of, the original population. From top to bottom, the process of choosing the degree of statistical noise to inject into the dataset, the process by which that noise is introduced, and the adjustments made afterward to comply with various constraints, is an exercise in statistical inference.

---

<sup>3</sup>Evans, Georgina, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta. “Statistically valid inferences from privacy protected data.” URL: [GaryKing.org/dp](https://garyking.org/dp) (2020).

## Step 1: Obtain accurate counts of people and geographies

The first step is to gather the actual numbers of people, their race, ethnicity, housing status, and geographic location. The Census Bureau notes that this step is largely accomplished via self-reports from individuals throughout the country with a significant amount of follow-up by Census workers.

## Step 2: Inject statistical noise

The second step is to inject a certain degree of statistical noise into the data. This process is referred to as “differential privacy”. There are a variety of approaches to differential privacy, and the proposed approach taken by the Census Bureau relies on basic statistical methods. At its core, differential privacy is an exercise in probability theory, and “[p]robability is the foundation and language for statistics.”<sup>4</sup> In describing differential privacy as a question of probability and statistical methods, Bambauer et. al (2013) state, “[D]ifferential privacy disclosure occurs when the probability that a query will return a particular result differs from the probability that a query would return that same result if the person were not included in the database. It also ensures that the inclusion of a person who isn’t in the dataset wouldn’t change the results of a query by too much. The measure of the disclosure for a particular query to a particular individual is the ratio of the probabilities that the query system would return the result with and without the individual’s data. Ideally, this ratio would be 1, allowing no disclosure at all. But since this is impossible to achieve if the responses are to be useful, the data curator can select some small level of disclosure that society is willing to tolerate. The closer to 1 the ratio is, the less disclosure has taken place.”<sup>5</sup> In other words, differential privacy is a process by which statistical noise is injected into the original data counts so as to obscure the true values in order to lower the probability

---

<sup>4</sup>Hwang, Jessica., Blitzstein, Joseph K.. Introduction to Probability, Second Edition. United States: CRC Press, 2019.

<sup>5</sup>Bambauer, Jane, Krishnamurthy Muralidhar, and Rathindra Sarathy. “Fool’s gold: an illustrated critique of differential privacy.” *Vand. J. Ent. & Tech. L.* 16 (2013): 701.

that an individual's identity and accurate information can be inferred from the data. The greater the noise, the lower this probability.

To determine the amount of noise to be injected into a particular quantity of interest (e.g. the number of men residing in a particular census block), the Census Bureau first determines the amount of overall privacy needed (the "privacy budget") and where to allocate that budget (for example how much to apply at the national, state, county, tract, block group, and block level). This budget is referred to by the greek letter epsilon. The size of epsilon determines the amount of statistical noise that is injected into the original, accurate counts.

In an interview with Science Magazine, John Abowd, chief scientist and associate director for research at the Census Bureau and Jerry Reiter, a professor of statistics at Duke University who has worked as a consultant with the Census Bureau discussed how epsilon is chosen. "Abowd says the privacy budget 'can be set at wherever the agency thinks is appropriate.' A low budget increases privacy with a corresponding loss of accuracy, whereas a high budget reveals more information with less protection. The mathematical parameter is called epsilon; Reiter likens setting epsilon to 'turning a knob.' And epsilon can be fine-tuned: Data deemed especially sensitive can receive more protection. The epsilon can be made public, along with the supporting equations on how it was calculated."<sup>6</sup> The Census Bureau has said, regarding the choice of epsilon, "Decisions about the privacy-loss budget (epsilon) for decennial products are made by a committee of senior career Census Bureau data experts, the Data Steward Executive Policy Committee (DSEP). The DSEP will analyze the results of internal and external research on the fitness-of-use of the 2010 Demonstration Data Products to make an informed decision on the level of epsilon for the 2020 Census data."<sup>7</sup> In other words, the Census Bureau is using information from the population and distribution of various demographics in the population to learn about and make statistical inferences regarding the total size of the privacy budget and the degree to which certain

---

<sup>6</sup><https://www.sciencemag.org/news/2019/01/can-set-equations-keep-us-census-data-private>

<sup>7</sup><https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products/faqs.html>

places and people's information needs more or less of that privacy budget allocated.

Once the privacy budget has been allocated, the Census Bureau will take draws from a probability distribution (think of a "draw" as akin to rolling a dice, but where the die has more than 6 sides and the probability of each side coming up is not equal) which will then be added to or subtracted from the accurate counts. The chosen value of epsilon has a direct relationship with the particular shape of the statistical distribution.

Probability distributions are a foundational tool upon which much of statistics operates. Using a probability distribution to inject statistical noise allows the researcher to be mathematically rigorous (as opposed to making ad hoc decisions about where and how much noise to inject) in describing the process by which the amount of statistical noise to be introduced is determined while simultaneously making it impossible for a person to reverse engineer the precise values by which counts are added to or subtracted from in any given case because draws from probability distributions are randomly determined. "Randomization is essential; more precisely, any non-trivial privacy guarantee that holds regardless of all present or even future sources of auxiliary information, including other databases, studies, Web sites, on-line communities, gossip, newspapers, government statistics, and so on, requires randomization."<sup>8</sup> Thus, the application of differential privacy can ultimately be considered as a particular application of probability, sampling, and statistics. The greater the statistical noise injected into the data, the lower the probability of a record linkage successfully occurring and privacy being revealed. Similarly, the smaller the statistical noise introduced into the data, the higher the probability of someone successfully identifying individuals included in the data.<sup>9</sup>

In the case of the 2020 Census, the Census Bureau has indicated that the particular probability distribution that they will use is the Laplace distribution, which is displayed in

---

<sup>8</sup>Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science* 9, no. 3-4 (2014): 211-407.

<sup>9</sup>See Dwork, Cynthia, and Adam Smith. "Differential privacy for statistics: What we know and what we want to learn." *Journal of Privacy and Confidentiality* 1, no. 2 (2010). for a technical discussion of the ideas presented in this paragraph.

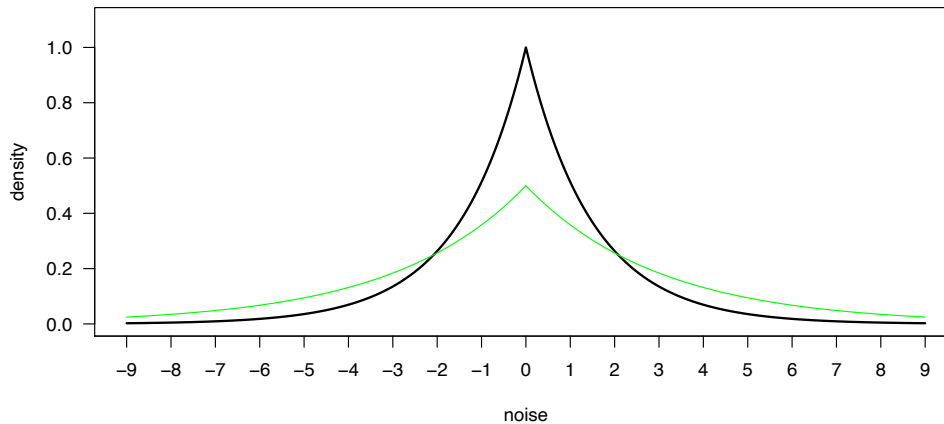


Figure 1: **Example Laplace Distribution** - The Laplace distribution is a symmetric probability distribution. The distribution can be steeper (black line) or flatter (green line) depending on the value set for the shape parameter. The higher the distribution on the vertical axis, the more likely are draws from the distribution (values on the x-axis) to have values in that region. For example, a draw of 1 is more likely than 5 or -3.

Figure 1 below.<sup>10</sup> The Laplace distribution is symmetric and rises to a single peak in the middle of the distribution. The higher the distribution on the vertical axis, the more likely are draws from the distribution to have values in that region. In other words, since the Laplace distribution is centered around zero, small numbers near zero are more likely to be drawn than are larger negative or positive numbers.

The particular spread of the Laplace distribution is determined by setting a shape parameter (epsilon), which can make the distribution more or less “flat.” The flatter the distribution (the green line in Figure 1), the more likely are draws to have larger values (either positive or negative) while a steeper distribution (the black line in Figure 1) is more likely to have draws with smaller values. In other words, the choice of the distribution’s spread (which is determined by the Census Bureau) injects more or less noise, on average, into the population counts depending on how “flat” the Census Bureau decides to make the Laplace distribution. Once these draws have been taken, the particular values are then

<sup>10</sup>In some cases the Census Bureau has indicated they use a two-sided geometric distribution, which is similar in shape to a Laplace distribution.



added to, or subtracted from, the original, accurate counts. Table 1 below shows a simplified example of this process in three steps. The top table shows the accurate counts of people in an area (such as a census block) based on their gender and race. The middle table then shows how statistical noise based on samples drawn from the Laplace distribution are added to or subtracted from the accurate counts.

The Department of Commerce’s reply report states that “Plaintiffs assert that differential privacy is a ‘statistical method’ — and perhaps it is in a colloquial sense — but the reasons they offer in support of that conclusion are untethered from the express statutory definition of ‘statistical method’ found in Section 209’s text (pg. 7).” While I have not been asked to speak to the relevance of differential privacy to Section 209, it is curious that the Department of Commerce refers to differential privacy as a statistical method “in a colloquial sense.” In fact, it is difficult to know what this even means. It is hard to imagine how differential privacy, which at its most basic level is adding or subtracting values sampled from a probability distribution function, could be seen as anything but an exercise in statistical methods. Probability, sampling, and the use of probability distributions, sit at the very foundations of statistics. It would be hard to find a statistics textbook that didn’t include a discussion of these ideas or that didn’t devote significant page space to the development of probability theory and probability distributions.<sup>11</sup>

---

<sup>11</sup>See for example:

Diez, David., Barr, Christopher., Çetinkaya-Rundel, Mine. *OpenIntro Statistics*. United States: OpenIntro, Incorporated, 2019.

Imai, Kosuke., Bougher, Lori D.. *Quantitative Social Science: An Introduction in Stata*. United States: Princeton University Press, 2021.

Hwang, Jessica., Blitzstein, Joseph K.. *Introduction to Probability, Second Edition*. United States: CRC Press, 2019.

all of which are used in introductory statistics courses at Harvard and Princeton Universities.

### **Step 3: Address fractions, negative numbers, and internal consistency**

In some cases the process of differential privacy would be complete after the researcher adds or subtracts from the original data the particular random draws that arose from sampling from the chosen statistical distribution. However, census data require several additional steps to address constraints that arise from the need for the differential privacy process to align with other objectives related to the use of census data. These steps are collectively referred to by the Census Bureau as “post-processing.”

The first issue centers on the need for counts of people, housing units, and other statistics to be reported as integers (as opposed to fractions). The middle panel of Table 1 illustrates this point. The particular draws from the Laplace distribution have been added to or subtracted from the original data. One problem is that there are now fractions of people living in this particular census block. To resolve this issue, fractional values are rounded to become integers. The second issue arises from cases in which the draw from the Laplace distribution subtracts more than the original number of people who occupy a particular cell in the table. This is especially likely to happen in cases with small counts of people, such as in census blocks. This results negative values, which are of course, not possible. To resolve this issue, these values are truncated so that they are no longer less than zero. A simplified example of this step is displayed in the bottom panel of Table 1.

The steps of integer rounding and resolving negative counts would be trivial except that the Census Bureau has committed to providing invariant (i.e. accurate) counts of people for redistricting purposes at the state level. However, when a block (or subgroup within a block) is rounded or adjusted to no longer be negative, this results in an overall change in the total population, as illustrated in Table 1 below. Thus, an equal number of people must be subtracted from another block (or subgroup within a block) to maintain the correct population numbers across the various states. Furthermore, a similar problem must be resolved with geographies that are nested within other geographies (i.e. the number of

people in all blocks in tract X should add up to the total population reported in tract X, even after the statistical noise has been added). Because noise is injected independently into each histogram, the totals are inconsistent with each other, both within and across geographic levels. This is an incredibly complex problem to solve since the number of ways in which blocks (or subgroups of blocks) could be adjusted to maintain the correct population at the state level while also making the data internally consistent across geographies is enormous.

Table 1: **A simplified example of differential privacy and post-processing:**

	Race			
	White	Black	Other	
Male	5	2	0	
Female	3	4	3	
Total Population:				17

*After adding statistical noise via sampling from probability distribution:*

	Race			
	White	Black	Other	
Male	$5+3=8$	$2+0=2$	$0-5=-5$	
Female	$3+2.25=5.25$	$1+.5=1.5$	$3-1=2$	
Total Population:				13.75

*After post-processing to remove fractions and negative counts:*

	Race			
	White	Black	Other	
Male	8	2	0	
Female	5	2	2	
Total Population:				19

To accomplish this objective, the Census Bureau uses what is known as a “least squares optimization,” which is another common statistical method. In this case, the problem to optimize over is incredibly large and unusually difficult given the size of the dataset as well as the numerous constraints imposed as a part of the optimization problem. Abowd et.

al (2020) provide a technical description of this least squares optimization problem.<sup>12</sup> The use of optimization via the method of least squares is an extremely common application of statistical inference and is widely used across the social sciences, natural sciences, and many other disciplines.<sup>13</sup>

In a presentation describing the differential privacy process and the 2020 Census, Michael Hawes, a Senior advisor for data access and privacy, described this post-processing optimization procedure as “statistical inference creating non-negative integer counts from the noisy measurements.”<sup>14</sup> In other words, the procedure is a particular use of statistical inference to locate the optimal solution (i.e. closest to the DP injected counts) to the problem of cell counts that need to be non-negative integers whose sum totals up to the accurate count of people at the state level, among other constraints. By their own admission, the Census Bureau is using statistical inferential methods to implement the post-processing procedure.

### 3 Enumeration and DP

In their report, the Department of Commerce appears to draw a hard distinction between the “enumeration” period of the census and the “disclosure avoidance” methods that are applied to the census data after they are enumerated. This distinction is, however, a matter of semantics and not one of substance. This is because, for all intents and purposes, users of the census data, including state legislatures and other redistricting bodies, will only have access to the noise-injected data and not the original, accurate, enumerated data. The Census Bureau argues that because they provide the accurate state-level data for the purposes of redistricting, that the differential privacy and post-processing procedure are not

<sup>12</sup><https://www2.census.gov/adrm/CED/Papers/CY20/202008AbowdBenedettoGarfinkelDahletal-The%20modernization%20of.pdf>

<sup>13</sup>These are only a small sample of statistics textbooks that discuss statistical inference and least squares methods in detail: Silvey, S.D.. *Statistical Inference*. Japan: Taylor & Francis, 1975. Berger, Roger L., Casella, George. *Statistical Inference*. United States: Cengage Learning, 2021. Stock, James H., Watson, Mark W.. *Introduction to Econometrics*. United States: Pearson Education, 2015. Freedman, David A.. *Statistical Models: Theory and Practice*. United States: Cambridge University Press, 2009. Greene, William H.. *Econometric analysis*. United Kingdom: Pearson/Prentice Hall, 2008.

<sup>14</sup><https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf>, slide 24

a part of the enumeration procedure. However, this is only partially true. In addition to the state level totals, redistricting bodies also need accurate counts at the sub-state level. While the state-level data are used to allocate seats for the US House of Representatives, the districts themselves require more fine-grained data to ensure equal population across districts within states as well as other racial and geographic-based measures to ensure the creation of certain majority-minority districts or the protection of other communities of interest, as required by law.

The adding and subtracting of counts that occurs during the differential privacy and post-processing stages of the enumeration process will impact the overall counts of people that are used to partition states into their various legislative districts. Importantly, in 2010 the Census Bureau provided accurate enumerations at both the state level and census block level, which allowed for not only an accurate allocation of legislative seats across the states, but also the accurate creation of legislative districts from the combination of census blocks within states. This will not be the case if the Census Bureau goes forward with their plan for differential privacy and post-processing of the data.

I, Michael Barber, am being compensated for my time in preparing this report at an hourly rate of \$400/hour. My compensation is in no way contingent on the conclusions reached as a result of my analysis.

A handwritten signature in black ink, appearing to read "Michael Barber". The signature is fluid and cursive, with the first name "Michael" and last name "Barber" clearly distinguishable.

Michael Barber

March 25, 2021

UNITED STATES DISTRICT COURT  
FOR THE MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION

THE STATE OF ALABAMA, et al., )  
)  
Plaintiffs, )  
)  
v. )  
)  
UNITED STATES DEPARTMENT OF )  
COMMERCE, et al., )  
)  
Defendants. )

Case No.:  
2:21-cv-00211-RAH-ECM-KCN

---

BRIEF OF AMICUS CURIAE PROFESSOR JANE BAMBAUER  
IN SUPPORT OF PLAINTIFFS'  
COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

---

**Christopher W. Weller**  
CAPELL & HOWARD, P.C.  
150 South Perry Street  
Montgomery, AL 3104  
Phone: (334) 241-8066  
Fax: (334) 241-8266  
chris.weller@chlaw.com

*Counsel for Amicus Curiae Professor Jane Bambauer*

**TABLE OF CONTENTS**

**TABLE OF CONTENTS** ..... i

**TABLE OF AUTHORITIES** ..... ii

**INTEREST OF *AMICUS CURIAE*** ..... 1

**SUMMARY OF ARGUMENT** ..... 1

**ARGUMENT** ..... 2

**I. Differential Privacy Uses a Flawed Conception of Privacy** ..... 2

**A. Differential Privacy Has No Relation to Real World Risk** ..... 3

**B. Differential Privacy Provides a False Sense of Precision and Certainty** ..... 10

**II. Traditional Disclosure Control Techniques Do a Better Job Protecting Privacy and Preserving Utility** ..... 12

**III. Neither Law Nor Public Distrust Can Justify the Census Bureau’s Decision to Adopt Differential Privacy** ..... 19

**A. Privacy Laws** ..... 19

**B. Public Trust** ..... 20

**IV. The Census Bureau’s Position Sets a Trap for Public Records Laws** ..... 22

**CONCLUSION** ..... 24

**CERTIFICATE OF SERVICE** ..... 26



**TABLE OF AUTHORITIES**

**CASES**

*ACLU Found. of Ariz. v. U.S. Dep’t Homeland Sec.*, No. CV-14-02052-TUC-RM (BPV),  
2017 WL 8895339 (D. AZ. Jan. 26, 2017) -----23

*ACLU v. Dep’t of Defense*, 543 F.3d 59 (2d Cir. 2008)-----24

*Brantley v. Kuntz*, 98 F. Supp. 3d 884 (W.D. Tex. 2015) -----18

*Floyd v. City of New York*, 959 F. Supp. 2d 540 (S.D. N.Y. 2013) -----23

*Motor Vehicle Mfrs. Ass’n of U.S., Inc. v. State Farm Mut. Auto. Ins. Co.*, 463 U.S. 29 (1983) -25

*St. Joseph Abbey v. Castille*, 712 F.3d 215 (5th Cir. 2013) -----18

*United States v. Carroll Towing Co.*, 159 F.2d 169 (2d Cir. 1947)----- 9

**OTHER AUTHORITIES**

2020 Disclosure Avoidance System Updates, U.S. CENSUS BUREAU  
<https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>----- 12

Cynthia Dwork, *A Firm Foundation for Private Data Analysis*,  
54 COMM’NS OF THE ACM 89 (2011) ----- 4

Daniel Kondor et al., *Towards Matching User Mobility Traces in Large-Scale Dataset*,  
IEEE Transactions on Big Data (Vol. 6, Issue 4) (Dec. 1, 2020) -----13

David Sidi & Jane Bambauer, *Plausible Deniability*,  
2020 PRIVACY IN STAT. DATABASES 91 (2020) -----12

David Van Riper, et al., *Differential Privacy and the Decennial Census*,  
IPUMS DIFFERENTIAL PRIVACY WORKSHOP (Aug. 15, 2019)  
[https://assets.ipums.org/files/ipums/intro\\_to\\_differential\\_privacy\\_IPUMS\\_workshop.pdf](https://assets.ipums.org/files/ipums/intro_to_differential_privacy_IPUMS_workshop.pdf)--16

Dept. Health & Human Servs., GUIDANCE REGARDING METHODS FOR DE-IDENTIFICATION OF  
PROTECTED HEALTH INFORMATION IN ACCORDANCE WITH THE HEALTH INSURANCE  
PORTABILITY AND ACCOUNTABILITY ACT (HIPAA) PRIVACY RULE (2012),  
<https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard> ----- 19

El Emam & Luk Arbuckle, ANONYMIZING HEALTH DATA: CASE STUDIES AND METHODS TO GET  
YOU STARTED 28 (2013) -----25

Federal Committee on Statistical Methodology,  
Statistical Policy Working Paper 22 (2d Version, 2005) -----12

Fida Kamal Dankar, *Estimating the Re-Identification Risk of Clinical Data Sets*,  
 12 BMC MED. INFORMATICS & DECISION MAKING 66 (2012)----- 13

Garret Christensen & Edward Miguel, *Transparency, Reproducibility, and the Credibility of  
 Economics Research*, 56 J. OF ECON. LITERATURE 920, 969 (2018)----- 6

Gina Kolata, *Your Data Were ‘Anonymized’? These Scientists Can Still Identify You*, N.Y. TIMES  
 (July 24, 2019), <https://www.nytimes.com/2019/07/23/health/data-privacy-protection.html>-- 13

Gregory E. Simon et al., *Assessing and Minimizing Re-Identification Risk in Research Data  
 Derived from Health Care Records*, 7 eGEMS 1, 3 (2019) ----- 13

Ian Lundberg, et al., *Privacy, Ethics, and Data Access: A Case Study of the Fragile Families  
 Challenge* (Sept. 10, 2019), <https://journals.sagepub.com/doi/10.1177/2378023118813023>-- 24

*In Massachusetts, laws intended to protect domestic abuse victims’ privacy are being used to  
 deny access to data about enforcement*, MUCKROCK (Jan. 9, 2018),  
<https://www.muckrock.com/news/archives/2018/jan/09/dv-mass-data/> ----- 24

James Lyall, et al., *Record of Abuse, Lawlessness and Impunity in Border Patrol’s Interior  
 Enforcement Operations*, AM. CIV. LIBERTIES UN. OF ARIZ., 4 (Oct. 2015) ----- 23

Jane Bambauer et al., *Fool’s Gold: An Illustrated Critique of Differential Privacy*,  
 16 VAND. J. ENT. & TECH. 727 (2014)----- 5

Jessie Gomez, *Louisiana judge grants access to state policy body-camera footage*,  
 MUCKROCK (Mar. 1, 2019) [https://www.muckrock.com/news/archives/2019/mar/01/louisiana-  
 bodycam/](https://www.muckrock.com/news/archives/2019/mar/01/louisiana-bodycam/) ----- 23

Josep Domingo-Ferrer & Krishnamurthy Muralidhar, *New Directions in Anonymization:  
 Permutation Paradigm, Verifiability by Subjects and Intruders, Transparency to Users*,  
 337 INFO. SCIS. 11, 12-13, 18 (2016) ----- 6

Joseph Neff, Ann Doss Helms, & David Raynor, *Why Have Thousands of Smart, Low-Income  
 NC Students Been Excluded from Advanced Classes?*, THE CHARLOTTE OBSERVER (May 21,  
 2017), <https://www.charlotteobserver.com/news/local/education/article150488822.html> ----- 24

Kathleen Benitez & Bradley Malin, *Evaluating re-identification risks with respect to the HIPAA  
 privacy rule*, 17(2) J. AM. MED. INFOR. ASS’N 169 (2010)----- 13

Kelsey Campbell-Dollaghan, *Sorry, Your Data Can Still Be Identified Even if It’s Anonymized*,  
 FAST COMPANY (Dec. 10, 2018), [https://www.fastcompany.com/90278465/sorry-your-data-  
 can-still-be-identified-even-its-anonymized](https://www.fastcompany.com/90278465/sorry-your-data-can-still-be-identified-even-its-anonymized)----- 13

Luc Rocher et al., *Estimating the Success of Re-Identifications in Incomplete Datasets Using  
 Generative Models*, 10 NATURE COMMS. art. 3069 (2019)----- 13

Mark Elliot & Josep Domingo-Ferrer, *The future of statistical disclosure control*, 3.1,  
 NAT’L STATISTICIAN’S QUALITY REV. INTO PRIVACY & DATA CONFIDENTIALITY METHODS  
 (2018) ----- 6, 24

Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*,  
 SCIENCE (Jan. 18, 2013) ----- 13

Michael B. Hawes, *U.S. Census Bureau, Implementing Differential Privacy: Seven Lessons From the 2020 United States Census*, HARV. DATA SCI. REV., Issue 2.2 (Apr. 30, 2020), <https://perma.cc/DB66-9B5R>-----22

Michael Hawes, *Differential Privacy and the 2020 Decennial Census*, U.S. CENSUS BUREAU (Jan. 28, 2020) presentation available at [https://zenodo.org/record/4122103/files/Privacy\\_webinar\\_1-28-2020.pdf](https://zenodo.org/record/4122103/files/Privacy_webinar_1-28-2020.pdf)----- passim

Natasha Singer, *With a Few Bits of Data, Researchers Identify ‘Anonymous’ People*, N.Y. TIMES BITS (Jan. 29, 2015, 2:01 PM), <https://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/>----- 13

Philip Leclerc, *The 2020 Decennial Census TopDown Disclosure Limitation Algorithm*, U.S. CENSUS BUREAU (Dec. 11, 2019), <https://www.nationalacademies.org/event/12-11-2019/docs/DCC854281ACE97996C107A2DC1BE711DFF02965EE0EC>----- 3

Ramachandran, *et al.*, *Exploring Re-identification Risks in Public Domains*, U.S. CENSUS BUREAU (Sept. 12, 2012) <https://www.census.gov/srd/papers/pdf/rrs2012-13.pdf>----- 14

Rebecca Jacobson, *Your ‘Anonymous’ Credit Card Data Is Not So Anonymous, Study Finds*, PBS NEWS HOUR (Jan. 29, 2015, 5:54 PM), <https://www.pbs.org/newshour/nation/anonymous-credit-card-data-anonymous-study-finds>----- 13

Sophie Bushwick, *‘Anonymous’ Data Won’t Protect Your Identity*, SCIENTIFIC AMERICAN (July 23, 2019), <https://www.scientificamerican.com/article/anonymous-data-wont-protect-your-identity/>----- 13

*Stop-And-Frisk 2011*, NEW YORK CIV. LIBERTIES UN. (May 2012) [https://www.nyclu.org/sites/default/files/publications/NYCLU\\_2011\\_Stop-and-Frisk\\_Report.pdf](https://www.nyclu.org/sites/default/files/publications/NYCLU_2011_Stop-and-Frisk_Report.pdf)-----23

*Stop-and-Frisk in the de Blasio era*, NEW YORK CIV. LIBERTIES UN. (Mar. 2019)-----23

Tapan K. Nayak et al., *Measuring Identification Risk in Microdata Release and Its Control by Post-Randomization*, CENTER FOR DISCLOSURE AVOIDANCE RESEARCH, U.S. CENSUS BUREAU ----- 6

Tennessee Watson, *Justice Isn’t Always Done for Child Sex Abuse-I Know Firsthand*, REVEAL (Aug. 11, 2016), <https://revealnews.org/article/tennessee-watson-justice-isnt-always-done-for-child-sexual-abuse-i-know-firsthand/>-----23

U.S. Census Bureau, *Why a Census?: How the Census Benefits Your Community*, <https://www.census.gov/programs-surveys/decennial-census/2020-census/about/why.html> --21

REGULATIONS

45 C.F.R. §169.103 -----19

STATUTES

13 U.S.C. § 181-----21

13 U.S.C. § 9-----20

### **INTEREST OF *AMICUS CURIAE***

Amicus is a Professor of Law at the University of Arizona and an expert in the public policy and industry practices related to privacy, research, and Big Data. Throughout my academic career, I have studied the societal risks and benefits related to the collection and use of personal data. Much of my scholarly and community service work relates to deidentified research data. In collaboration with statistical disclosure experts, I have written guidance documents, scholarly publications, and an amici curiae brief for the U.S. Supreme Court. I have worked with the ACLU of Arizona to facilitate public access to deidentified data on Border Patrol detainees. I have served on the Program Committee for UNESCO's annual conference on Privacy in Statistical Databases, and I have given presentations about the trade-off between privacy risk and research to the U.N. Economic Commission for Europe/Eurostat, the Federal Trade Commission, and Google.

I have no personal interest in the outcome of this case, but a professional interest concerning the impact that the adoption of Differential Privacy could have on government accountability and open research. As the government and private companies have access to increasing amounts of personally identifiable information, it is more important than ever that researchers, nonprofits, and journalists have access to accurate statistical data.

### **SUMMARY OF ARGUMENT**

The State of Alabama has done an excellent job illustrating how the Census Bureau's use of Differential Privacy will affect the accuracy and reliability of nearly every statistical table and data product that is in use for highly consequential redistricting and resource allocation decisions. This amicus brief contributes a more fundamental critique and objection to Differential Privacy as a tool for mitigating risk in public datasets.

The Census Bureau’s adoption of Differential Privacy is indefensible because the definition and measure of “privacy” imbedded in Differential Privacy is poorly matched to actual risk of disclosure. Because “privacy” is defined in a manner that is insensitive to context, including which types data are most vulnerable to attack, Differential Privacy compels data producers to make bad and unnecessary tradeoffs between utility and privacy. Reidentification attacks that are much more feasible, and thus much more likely to occur, are treated exactly the same as absurdly unlikely attacks. As a result, whatever “privacy” budget is chosen, the resulting noise-added data is simultaneously less accurate *and less privacy-protective* than a traditional disclosure control method that is attuned to context.

Thus, there is no rational basis for employing Differential Privacy. Differential Privacy, if used as intended, would wreak havoc on the accuracy of almost all US Census data products and defeat the very purpose for comprehensive Census data collection without any meaningful gain in the (already adequate) privacy protections. And it is particularly irrational given that the delays caused by implementing Differential Privacy will have serious consequences for elections this year. For these reasons, the adoption of Differential Privacy is an arbitrary and capricious use of the agency’s discretion to balance competing societal interests in statistical accuracy and data privacy.

## **ARGUMENT**

### **I. Differential Privacy Uses a Flawed Conception of Privacy**

Differential Privacy guarantees to each data subject that the probability a statistical report will present a particular value is not too different from the probability that it would give the same value even if the data subject wasn’t included in the dataset. As a practical matter, the guarantee requires a certain amount of noise (*i.e.*, the intentional introduction of precisely calibrated error)

to be added, and the amount of noise is determined by a worst-case scenario in which an attacker might know everything about a database except one last detail.

The Census Bureau chose to adopt Differential Privacy rather than continuing to use traditional disclosure control methods for two key reasons: Differential Privacy makes no assumptions about the reidentification attacks that could be possible now or in the future; and it quantifies the concept of “privacy” in a way that allows the Bureau to make and meet certain guarantees. However, each of these purported advantages of Differential Privacy is in fact detrimental to the Census Bureau’s mission.

**A. Differential Privacy Has No Relation to Real World Risk**

Because Differential Privacy measures privacy under worst case scenarios, the privacy protections that are guaranteed by Differential Privacy are not dependent on context. No data steward has to make predictions or value judgments about which types of data are more vulnerable to reidentification attack, and which types are more sensitive and harmful if discovered. As the Census Bureau itself explains, Differential Privacy “does not directly measure re-identification risk (which requires specification of an attacker model). Instead, it defines the maximum privacy “leakage” of each release of information compared to some counterfactual benchmark (*e.g.*, compared to a world in which a respondent does not participate, or provides incorrect information.)”<sup>1</sup>

---

<sup>1</sup> Philip Leclerc, *The 2020 Decennial Census TopDown Disclosure Limitation Algorithm*, U.S. CENSUS BUREAU (Dec. 11, 2019) presentation available at <https://www.nationalacademies.org/event/12-11-2019/docs/DCC854281ACE97996C107A2DC1BE711DFF02965EE0EC> (last accessed Apr. 6, 2021).

This is characterized as a benefit by the Census Bureau as well as computer scientists who developed Differential Privacy since it automatically guards against every conceivable or hypothetical attack.<sup>2</sup>

**Differential Privacy**

aka “Formal Privacy”

- quantifies the precise amount of privacy risk...
- for all calculations/tables/data products produced...
- no matter what external data is available...
- now, or at any point in the future!

18

Shape your future  
START HERE >

United States  
Census  
2020

However, the indifference to context is actually a drawback if the goal is to mitigate real world risk. The differential privacy model treats all data leakage the same, and all possible attacks as equally plausible. This is because privacy loss is measured based on an intruder who knows *everything* about *every person* except for one last piece information about one person.<sup>3</sup> Because the context-free definition of privacy leakage is so easily triggered, the privacy “guarantees” offered by Differential Privacy are deceptive. After all, in order to produce any useful data, the data steward must allow for *some* potential information leakage. The data steward does this by

---

<sup>2</sup> Michael Hawes, *Differential Privacy and the 2020 Decennial Census*, U.S. CENSUS BUREAU (Jan. 28, 2020) presentation available at [https://zenodo.org/record/4122103/files/Privacy\\_webinar\\_1-28-2020.pdf](https://zenodo.org/record/4122103/files/Privacy_webinar_1-28-2020.pdf) (last accessed Apr. 6, 2021) (hereinafter “Hawes presentation”).

<sup>3</sup> Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMM’NS OF THE ACM 89, 92 (2011).

selecting parameters like  $\epsilon$  (“epsilon”), which allows some statistical data to be produced as long as the reports put a limit on the confidence that a nearly-omniscient attacker would have when receiving new information.<sup>4</sup> But these parameters do not and cannot ensure that the relaxations in privacy are well-aligned with real world risk. That is, if  $\epsilon$  is large so as to allow reasonable levels of accuracy, it is just as likely to be “spent” on statistical products that we *know* are vulnerable to reidentification attack as it is on products that we have good reason to believe is not likely to be reidentified.

For example, when constructing the limited types of data that are available in enumeration district files, Differential Privacy requires the Census Bureau to protect against privacy leakage pertaining to Hispanic status. What this means is that noise must be added to thwart a hypothetical intruder who has access to the race, age, Census block, and housing type of a particular target as well as the race, age, Census block, housing type, and Hispanic status of *every single other person* in the target’s district because this hypothetical intruder might then use the Census file to determine the Hispanic status of the target. The Census Bureau can of course spend some of its privacy budget to allow for more accurate reporting of data, but this privacy budget expenditure is wasteful. Nobody now or in the future will have access to that much auxiliary information in a form that reports exactly the same values as the Census data, and if they did, it’s hard to believe they wouldn’t know the Hispanic status of that last person. Yet by spending any part of a privacy budget to guard against this figment of the imagination, some other data table of high consequence will have to be made less accurate. If traditional disclosure control techniques can be criticized for

---

<sup>4</sup> For an illustrated explanation of Differential Privacy and the meaning of epsilon, see Jane Bambauer et al., *Fool’s Gold: An Illustrated Critique of Differential Privacy*, 16 VAND. J. ENT. & TECH. 727 (2014).



failing to anticipate some types of attacks, Differential Privacy can be criticized for anticipating *all* of them.

A formal assumption that attackers will be virtually omniscient makes privacy protection easier for the Census Bureau because it relieves the agency from having to make educated (but uncertain) predictions about which types of threats are plausible and which are not. The adoption of Differential Privacy therefore shields the Census Bureau from criticism that the agency made errors in judging which threats were more or less plausible. But the same formalism that is held up as a benefit of Differential Privacy permits an abdication of the responsibility to assess risks realistically, and to use mitigating strategies (like the addition of noise) where they are most needed.<sup>5</sup>

Consider, for example, what would have happened if the Department of Health and Human Services had decided to implement Differential Privacy when it produced public data on COVID cases and hospitalizations. Even if data tables were produced only one a week (instead of daily) in

---

<sup>5</sup> Josep Domingo-Ferrer & Krishnamurty Muralidhar, *New Directions in Anonymization: Permutation Paradigm, Verifiability by Subjects and Intruders, Transparency to Users*, 337 INFO. SCIS. 11, 12-13, 18 (2016); Tapan K. Nayak et al., *Measuring Identification Risk in Microdata Release and Its Control by Post-Randomization*, CENTER FOR DISCLOSURE AVOIDANCE RESEARCH, U.S. CENSUS BUREAU (assessing the problem with formal privacy measures, like “differential privacy,” and concluding “[t]hus, for developing practical disclosure control goals, it is essential for the agency to consider intruders with limited prior information about their target units.”); Mark Elliot & Josep Domingo-Ferrer, *The future of statistical disclosure control*, 3.1, NAT’L STATISTICIAN’S QUALITY REV. INTO PRIVACY & DATA CONFIDENTIALITY METHODS (2018) (“Many authors have commented that this environment is inherently difficult—if not impossible—to understand and therefore directly assessing risk is itself impossible. This in turn has led to bad decision-making about data sharing (a strange mixture of over-caution and imprudence which is driven more often than not by the personality of the decision-maker rather than by rational processes.)”); Garret Christensen & Edward Miguel, *Transparency, Reproducibility, and the Credibility of Economics Research*, 56 J. OF ECON. LITERATURE 920, 969 (2018) (“They have established that there is inherently a trade-off between these two objectives (Dwork and Smith 2010; Heffetz and Ligett 2014), though few actionable approaches to squaring this circle are currently available to applied researchers, to our knowledge.”).

order to preserve the “privacy budget,” a year’s worth of data on current hospitalizations and weekly case numbers would cause the data to be useless. Here, for example, is what a table of case counts and hospitalizations might look like for a sample of Alabama counties if the tables were produced using an epsilon of one (assuming that the department produces weekly tables, and *does not* produce any other data.)

**Example of Differential Privacy Applied to COVID Data (epsilon = 1)**

County	Case #s this week		14-Day Change		Currently Hospitalized		14-Day Change	
	True	With DP	True	With DP	True	With DP	True	With DP
Jefferson ›	420	237	-35%	-68%	157	146	8%	-51%
Madison ›	196	626	-34%	Infinite	62	0	-19%	0%
Montgomery ›	175	260	-11%	2500%	47	54	0%	32%
Tuscaloosa ›	168	126	-39%	-75%	21	215	-16%	Infinite
Mobile ›	140	215	-61%	41%	16	0	-62%	-100%
Shelby ›	140	253	-37%	26%	158	136	6%	27%
Baldwin ›	84	452	-47%	Infinite	56	197	-35%	Infinite
Lee ›	70	183	-31%	151%	9	0	-25%	0%
Talladega ›	63	118	-23%	Infinite	150	101	8%	-20%
Elmore ›	63	94	-55%	-58%	52	1	-13%	Infinite
Lauderdale ›	56	58	-18%	Infinite	5	41	0%	105%
Cullman ›	56	51	-29%	-50%	5	0	67%	0%
St. Clair ›	49	83	-9%	-48%	159	208	6%	-60%
Calhoun ›	49	264	-25%	Infinite	16	49	-30%	Infinite
Autauga ›	49	0	0%	-100%	65	102	-6%	Infinite
Marshall ›	49	55	17%	-53%	64	331	-19%	Infinite
Limestone ›	49	60	9%	-15%	65	9	-17%	-80%
Houston ›	49	0	40%	-100%	26	0	-13%	-100%
Chilton ›	35	0	-3%	-100%	60	265	0%	15%
Blount ›	35	0	0%	0%	148	227	9%	11%
Tallapoosa ›	35	0	13%	0%	9	0	-25%	-100%
Walker ›	35	97	-33%	62%	154	278	10%	Infinite
Morgan ›	28	86	-45%	-5%	75	0	-19%	-100%
Colbert ›	28	168	-15%	Infinite	7	0	-36%	0%
Etowah ›	28	0	-64%	-100%	13	13	-19%	Infinite
Jackson ›	21	0	-63%	-100%	135	227	2%	-15%
Russell ›	21	36	-68%	Infinite	40	46	-33%	-88%
Marion ›	14	0	-56%	-100%	4	86	100%	Infinite
Dale ›	14	12	-30%	-91%	29	122	-19%	Infinite
Coffee ›	14	0	27%	-100%	0	0	-100%	0%

Data sourced by the New York Times from the U.S. Department of Health & Human Services and state and local public health departments.

Even with a very generous “privacy budget” of 16 (the largest the U.S. Census Bureau has analyzed from its study of 2010 decennial data), several counties would miss critical trends or

harbor false senses of security and threat. And again, these tables assume that *no other data* will be reported. If the results were broken down by age, gender, or race, the error would be much worse.

**Example of Differential Privacy Applied to COVID Data (epsilon = 16)**

County	Case # this week		14-Day Change		Currently Hospitalized		14-Day Change	
	True	With DP	True	With DP	True	With DP	True	With DP
Jefferson ›	420	417	-35%	-36%	157	180	8%	28%
Madison ›	196	200	-34%	-37%	62	68	-19%	-14%
Montgomery ›	175	174	-11%	-13%	47	51	0%	24%
Tuscaloosa ›	168	165	-39%	-38%	21	20	-16%	-17%
Mobile ›	140	151	-61%	-58%	16	0	-62%	-100%
Shelby ›	140	156	-37%	-29%	158	150	6%	0%
Baldwin ›	84	85	-47%	-49%	56	65	-35%	-20%
Lee ›	70	68	-31%	-31%	9	0	-25%	-100%
Talladega ›	63	66	-23%	-23%	150	157	8%	8%
Elmore ›	63	71	-55%	-46%	52	37	-13%	-41%
Lauderdale ›	56	56	-18%	-19%	5	2	0%	-75%
Cullman ›	56	59	-29%	-20%	5	7	67%	-22%
St. Clair ›	49	36	-9%	-12%	159	164	6%	11%
Calhoun ›	49	47	-25%	-33%	16	25	-30%	0%
Autauga ›	49	49	0%	32%	65	65	-6%	-6%
Marshall ›	49	65	17%	44%	64	60	-19%	-13%
Limestone ›	49	36	9%	3500%	65	69	-17%	-8%
Houston ›	49	50	40%	-11%	26	26	-13%	-16%
Chilton ›	35	24	-3%	100%	60	63	0%	80%
Blount ›	35	16	0%	-67%	148	147	9%	2%
Tallapoosa ›	35	36	13%	44%	9	10	-25%	-9%
Walker ›	35	50	-33%	-2%	154	158	10%	36%
Morgan ›	28	20	-45%	-64%	75	78	-19%	-10%
Colbert ›	28	20	-15%	-13%	7	0	-36%	-100%
Etowah ›	28	22	-64%	-69%	13	20	-19%	33%
Jackson ›	21	17	-63%	-74%	135	145	2%	31%
Russell ›	21	20	-68%	-62%	40	33	-33%	-50%
Marion ›	14	11	-56%	-69%	4	0	100%	0%
Dale ›	14	15	-30%	-12%	29	43	-19%	19%
Coffee ›	14	18	27%	29%	0	0	-100%	-100%

The reason so much noise must be added to these tables in order to satisfy differential privacy is because the tables must be robust from an attack by a person who knows *every single person's COVID status in a given county except one person's* (the target's). Traditional methods of disclosure control would not make this preposterous assumption. Instead, with this limited data

(county-level geographic units, and no demographic data included), very little noise would be added, and that noise would focus on less populous counties or counties with a small number of hospitals and testing facilities.

One can analogize the Differential Privacy standard for privacy guarantees to the standards that courts had to develop in negligence cases to see the problem. The Census Bureau had been using statistical disclosure techniques that are consistent with the Hand formula from *United States v. Carroll Towing Co.*, 159 F.2d 169 (2d Cir. 1947). Risk was estimated based on the probability (p) that a misfeasor would have the auxiliary information to launch a successful attack and the losses (L) that would result from the disclosure of sensitive information. Risk under traditional notions of reasonableness would account for remote risks as well as common ones—threats that are very unlikely to materialize as well as those that are more common. But all would be appropriately weighted to reflect the probability and harm.

In contrast, by adopting Differential Privacy, the Census Bureau limits the public access and utility of Census data based on the worst-case hypotheticals. Differential Privacy guarantees are deliberately indifferent to real world considerations of risk. Differential Privacy defines privacy loss not based on what is foreseeable, but based on *the full universe* of hypotheticals. In the torts context, it would be equivalent to asking “if an omnipotent and all-powerful alien entered the scene, what could go wrong?” Indeed, the Census Bureau’s own explanation of their definition of privacy risk assumes that an attacker “has infinite computing resources, infinitely powerful algorithms, and allows her to have arbitrary side knowledge.”<sup>6</sup>

---

<sup>6</sup> Hawes presentation, *supra* note 2.

Arbitrary is the right word. The decision of the Census Bureau to abandon a standard of privacy protection based on foreseeable risks and to instead use a standard driven by nightmare fantasies is arbitrary and capricious, and an abuse of the Census Bureau’s discretion.

**B. Differential Privacy Provides a False Sense of Precision and Certainty**

A second benefit of Differential Privacy, which is related to the first, is that it allows privacy to be measured without the error or uncertainty that comes with predicting which reidentification attacks are more or less feasible. It measures privacy loss in a theoretical sense, with mathematical certainty, rather than in an actuarial sense. The Census Bureau claims that Differential Privacy’s ability to measure with certainty makes it “substantially better” than traditional methods for protecting privacy.<sup>7</sup>

**Comparing Methods**

**Data Accuracy**  
Differential Privacy is not inherently better or worse than traditional disclosure avoidance methods.  
Both can have varying degrees of impact on data quality depending on the parameters selected and the methods’ implementation.

**Privacy**  
Differential Privacy is substantially better than traditional methods for protecting privacy, insofar as it actually allows for measurement of the privacy risk.

26

Shape your future  
START HERE >

United States  
Census  
2020

However, the precision and certainty of Differential Privacy’s measure of “privacy risk” is only valuable for measuring risk from theoretical worse case scenarios. In other words, the word

---

<sup>7</sup> Hawes presentation, *supra* note 2.

“privacy” in the Differential Privacy literature is a term of art that means something specific, and that does not account for probability. If the Census Bureau desired instead to measure risk with some attunement to the probability that an attack could be attempted or could succeed, Differential Privacy is inferior to the traditional methods that model different threat scenarios and quantify risks under a range of assumptions.

Thus, rather than being beneficial, the quantitative precision of Differential Privacy is actually a drawback. Differential Privacy has the patina of mathematical elegance without actually quantifying privacy risks of the sort that most people care about. Indeed, when I explain the meaning of privacy risk (or privacy loss) to lay audiences, people often respond that the privacy budget should depend on whether the variables disclosed in the statistical data are more vulnerable (large “ $p$ ”, in the Hand formula sense) or sensitive (large “ $L$ ”). This, of course, is a reinvention of the disclosure avoidance techniques that the Census Bureau has used in the past and has now rejected with the adoption of Differential Privacy. Thus, it is useful to distinguish Differential Privacy’s concept of abstract privacy loss from privacy *risks* based on probability and harm.

The precision of Differential Privacy’s definition of “privacy” loss and its promise of privacy “guarantees” is also deceptive. Realistically, as the State of Alabama and its experts have shown, Census Bureau data cannot be produced in any useful form without using fairly generous values of the parameter  $\epsilon$ . Indeed, when the Census Bureau produced an exemplary file of 2010 Census data with Differential Privacy techniques, the Bureau explained that it set a more “conservative” privacy-loss budget than it expects will be set for the 2020 census—meaning that the demonstration data had “more noise (error) than should be expected in the final 2020 Census

data products[.]”<sup>8</sup>This is welcome news for those who are concerned about accuracy, but the implicit result is that more accuracy will come at the cost of greater privacy loss. These losses may be trivial or they may be very risky under real world conditions. Differential Privacy does not distinguish between these.

## **II. Traditional Disclosure Control Techniques Do a Better Job Protecting Privacy and Preserving Utility**

In the past, the U.S. Census Bureau successfully managed the risks inherent to public data releases using a range of disclosure control techniques. These methods often require data stewards to anticipate the most likely threats to data subjects, identify the most vulnerable records, and reduce the vulnerability with an eye toward preserving research potential. These techniques include data swapping, sampling, and blank-and-impute procedures that add uncertainty and error to the variables that are potentially vulnerable to reidentification attack.<sup>9</sup> Disclosure control is a highly pragmatic exercise that requires some grounded predictions of current and future behavior in order to make sure that the noise added to a dataset is strategically placed where a misfeasor is likely to attack. Privacy risk using these techniques is quantifiable, but requires some assumptions to be made about which attacks are remotely plausible and which are not.<sup>10</sup>

These techniques are not broken. Public use research datasets have continued to be safely produced without evidence of significant risk or harm to research subjects. Although there are

---

<sup>8</sup> 2020 Disclosure Avoidance System Updates, U.S. CENSUS BUREAU <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html> (last accessed Apr. 6, 2021).

<sup>9</sup> Federal Committee on Statistical Methodology, Statistical Policy Working Paper 22 (2d Version, 2005).

<sup>10</sup> For a description of various methods to quantify privacy risk outside Differential Privacy, see David Sidi & Jane Bambauer, *Plausible Deniability*, 2020 PRIVACY IN STAT. DATABASES 91 (2020).

many studies and popular media reports that state deidentified data can be easily reidentified<sup>11</sup>, the underlying research often relies on uniqueness of a data subject as the measure of reidentification risk, and simply assume attackers will possess ample information about their targets in identified form.<sup>12</sup> Moreover, even when an attacker *does* have significant amounts of auxiliary data, attacks are often so riddled with error that reidentifications are more likely to be wrong than right.

---

<sup>11</sup> Gina Kolata, *Your Data Were ‘Anonymized’? These Scientists Can Still Identify You*, N.Y. TIMES (July 24, 2019), <https://www.nytimes.com/2019/07/23/health/data-privacy-protection.html>; Sophie Bushwick, *‘Anonymous’ Data Won’t Protect Your Identity*, SCIENTIFIC AMERICAN (July 23, 2019), <https://www.scientificamerican.com/article/anonymous-data-wont-protect-your-identity/>; Kelsey Campbell-Dollaghan, *Sorry, Your Data Can Still Be Identified Even if It’s Anonymized*, FAST COMPANY (Dec. 10, 2018), <https://www.fastcompany.com/90278465/sorry-your-data-can-still-be-identified-even-its-anonymized>; Rebecca Jacobson, *Your ‘Anonymous’ Credit Card Data Is Not So Anonymous, Study Finds*, PBS NEWS HOUR (Jan. 29, 2015, 5:54 PM), <https://www.pbs.org/newshour/nation/anonymous-credit-card-data-anonymous-study-finds>; Natasha Singer, *With a Few Bits of Data, Researchers Identify ‘Anonymous’ People*, N.Y. TIMES BITS (Jan. 29, 2015, 2:01 PM), <https://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/>.

<sup>12</sup> Daniel Kondor et al., *Towards Matching User Mobility Traces in Large-Scale Dataset*, IEEE Transactions on Big Data (Vol. 6, Issue 4) (Dec. 1, 2020) (assessing “matchability” rather than reidentifiability, and finding an attacker could match 17% of the data subjects using “only” one week of comprehensive mobility information); Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, SCIENCE (Jan. 18, 2013) (concluding that intruders who already have the DNA sequence of a male relative might be able to identify a person in a genomic research database). For a critical take on using uniqueness as reidentification, see Gregory E. Simon et al., *Assessing and Minimizing Re-Identification Risk in Research Data Derived from Health Care Records*, 7 eGEMS 1, 3 (2019) (“To use a financial analogy, the exact amount (in dollars and cents) of the last 5 transactions in any credit account may be unique, but it would only be identifying to an adversary who already had access to those banking records.”); Fida Kamal Dankar, *Estimating the Re-Identification Risk of Clinical Data Sets*, 12 BMC MED. INFORMATICS & DECISION MAKING 66 (2012); Luc Rocher et al., *Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models*, 10 NATURE COMMS. art. 3069 (2019); Kathleen Benitez & Bradley Malin, *Evaluating re-identification risks with respect to the HIPAA privacy rule*, 17(2) J. AM. MED. INFOR. ASS’N 169 (2010) (“If a researcher receives a dataset drawn at random from the population of Ohio under Limited Dataset provisions, more than 1 out of 6 of those represented would be unique based on demographic information. Remember, though, that uniqueness is not sufficient to claim re-identification. There is still need for an identified dataset and VOTER reflects this reality. While higher than the risk under Safe Harbor, <LIMITED, VOTER> is significantly lower than <LIMITED, GENERAL>, particularly for smaller values of *g*. According to <LIMITED, VOTER>, only 0.002% of the population is 1-distinct and 0.01% is 5-distinct.”)



Internal studies performed by the U.S. Census Bureau to test their past uses of disclosure control techniques demonstrate that traditional disclosure control techniques provide excellent protection against reidentification attacks. For example, in one study, a group of Census researchers attempted to attack the data from an individual-level public use dataset on over two million data subjects. The data subjects were selected from three counties that were specifically chosen because of their vulnerability (residents in these counties are less transient, and therefore less likely to have noisy or stale data.) Next, the researchers purchased identified data on 700,000 people in the selected counties from a data aggregator and used all available overlapping key variables such as age, ethnicity, gender, and income. Out of the more than 2 million records in the research data files, the researchers' matching algorithm made apparent matches on 389 individuals. However, of those 389 apparent matches, *only 87 were actually correct*—an accuracy rate of just 22%.<sup>13</sup> Most of the apparent matches were wrong.

The Census Bureau's more recent examination of the 2010 census records found greater numbers of apparent matches, but the attempted attacks were similarly lousy in making accurate matches. This time, the Census Bureau used all 309 million U.S. census records and used census block, sex, and age to match census records to a commercially available database. This time, the researchers were able to make matches on 45% of the records (a whopping 138 million individuals), presumably because of the value that block-level geographic area provides for making unique matches. However, the vast majority of those matches (62%) were *wrong*.<sup>14</sup>

---

<sup>13</sup> Ramachandran, *et al.*, *Exploring Re-identification Risks in Public Domains*, U.S. CENSUS BUREAU (Sept. 12, 2012) accessible via <https://www.census.gov/srd/papers/pdf/rrs2012-13.pdf>.

<sup>14</sup> Hawes presentation, *supra* note 2.

## Reconstructing the 2010 Census: What Did We Find?

1. On the 309 million reconstructed records, census block and voting age (18+) were correctly reconstructed for all records and for all 6,207,027 inhabited blocks.
2. Block, sex, age (in years), race (OMB 63 categories), and ethnicity were reconstructed:
  1. Exactly for 46% of the population (142 million individuals)
  2. Within +/- one year for 71% of the population (219 million individuals)
3. Block, sex, and age were then linked to commercial data, which provided putative re-identification of 45% of the population (138 million individuals).
4. Name, block, sex, age, race, ethnicity were then compared to the confidential data, which yielded confirmed re-identifications for 38% of the putative re-identifications (52 million individuals).
5. For the confirmed re-identifications, race and ethnicity are learned correctly, though the attacker may still have uncertainty.

16 2020CENSUS.GOV

Shape  
your future  
START HERE >

United States  
Census  
2020

The Census Bureau presents this internal study as evidence that the Bureau needs to abandon traditional privacy methods and use Differential Privacy for the 2020 Census, but logic is strained. First, the conclusion that “the attacker may still have some uncertainty” is a dramatic understatement. It is misleading to suggest that 52 million individuals were accurately reidentified when the simulated attacker would not be able to distinguish them from the other 86 million individuals that the attacker falsely reidentified. The big numbers of reidentifications are meaningless if the Census Bureau credibly shows that even attacks that seem to succeed are most likely to be wrong.

Moreover, when the Census Bureau applied the same simulation attack methods on data that it had prepared with Differential Privacy standards, confirmed reidentifications were in the same ballpark (about 25 million accurate reidentifications for an epsilon value of 16.) Thus, the advantages of switching to Differential Privacy are modest.<sup>15</sup>

---

<sup>15</sup> Hawes presentation, *supra* note 2.

## Impact on Privacy

Using exactly the same re-identification strategy, we analyzed the differentially private microdata for persons at different privacy-loss budgets from  $\epsilon=0$  to  $\epsilon=16$ .

We used  $\epsilon=4$  for the differentially private person-level microdata computed for the 2010 Demonstration Data Products.

Results varied from a confirmed re-identification rate of 0 at  $\epsilon=0$  to 8.2% at  $\epsilon=16$ .

35 2020CENSUS.GOV

CBDRB-FY20-103

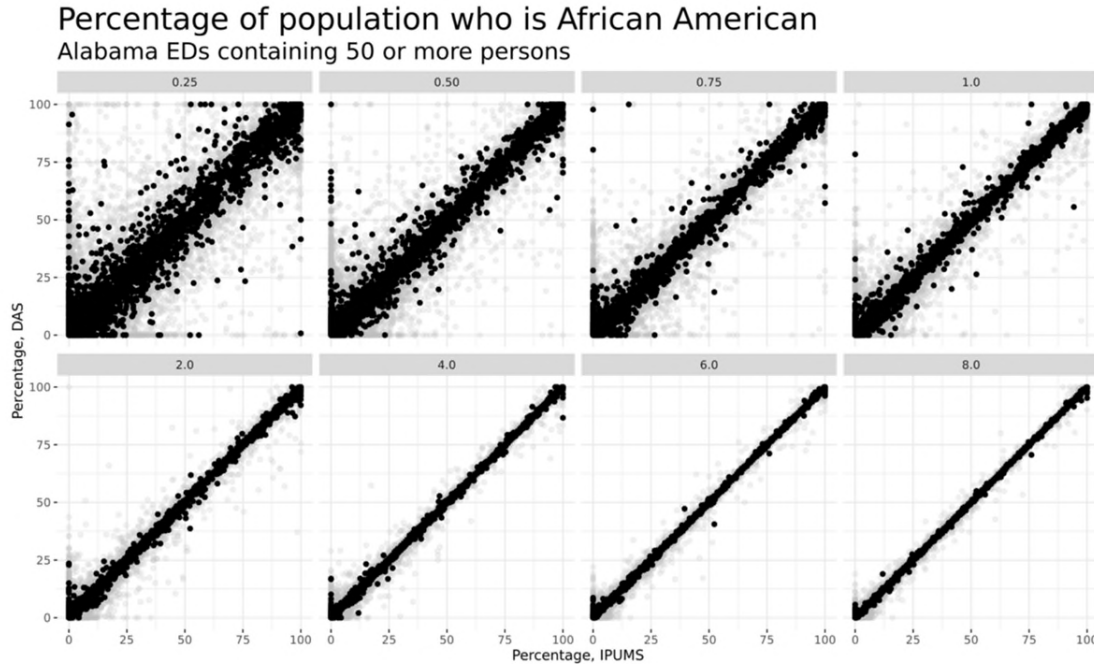
Shape  
your future  
START HERE >

United States  
Census  
2020

By contrast, the disadvantages from Differential Privacy, in terms of utility loss, are severe. The State of Alabama and its expert witnesses have highlighted some of the difficulties that will arise from needless inaccuracies in the Census data, including the likelihood that flawed data will cause redistricting errors. Pls' Mot. for a Prelim. Injun., Doc. No. 3 at 35. The same flawed data will deprive nonprofit organizations of the opportunity to investigate or challenge voting rights violations, too. Given that the error in Black/African American residency can be off by hundreds in many of Alabama's key legislative districts (*Id.*), illegal redistricting would be hard to even allege, let alone prove. Even if the Census Bureau uses a larger "privacy budget" (epsilon of 6 or 8), we can still expect noise to cause minority representation to be over- or under-reported in a few districts, as David Van Riper, *et al.* demonstrated in their study of the 1940 differentially private decennial census files.<sup>16</sup>

---

<sup>16</sup> David Van Riper, et al., *Differential Privacy and the Decennial Census*, IPUMS DIFFERENTIAL PRIVACY WORKSHOP (Aug. 15, 2019) available at [https://assets.ipums.org/\\_files/ipums/intro\\_to\\_differential\\_privacy\\_IPUMS\\_workshop.pdf](https://assets.ipums.org/_files/ipums/intro_to_differential_privacy_IPUMS_workshop.pdf).



These problems are bedeviling even for the relatively simple decennial census files, which do not report rich information about income and other sensitive, important traits. Once the Census Bureau begins to implement Differential Privacy to produce tables on income bands broken out by race, gender, and region, even a generous “privacy budget” will be spread so thin that the tables will become gibberish. Consider, again, the Alabama COVID tables presented above. If public health officials were to report the same figures broken down by age categories, race, and gender, the results would become even more erroneous. If the data is further segmented into smaller geographic or social units in order to understand whether, e.g., schools or nursing homes are having an outbreak, Differential Privacy would either prevent any meaningful statistics to emerge, or would require data stewards to select such a large “privacy budget” that realistic risks are unguarded.

The communities most likely to suffer from both the unjustified error and the unnecessary tolerance of (real world) privacy risk are small or vulnerable ones. Given that context-aware

assessments of privacy risk can outperform Differential Privacy, both for data utility and for protection from foreseeable threats, the Census Bureau’s decision to use Differential Privacy is an unreasonable use of agency discretion.

By contrast, traditional disclosure control experts would add just enough noise to the table cells of counties with very small numbers of cases or with only a few sources of treatment and testing to cause error and uncertainty for the remotely plausible attacks in which a neighbor or doctor might already know nearly everybody who has tested positive for COVID. More noise or error would be introduced for tables that report on commonly known demographics or characteristics (such as race or status as a student) since an attacker could plausibly know the demographics and basic characteristics of the relevant population. But traditional disclosure control techniques would not have to anticipate that an attacker, say, knows the current and past COVID status of every individual in a county except one (or, possibly, even knows that last person’s COVID status but does not know that target’s race.) These attack scenarios, however, are treated as just as likely as any other. This is why so much noise must be added to tables of simple counts in order to meet the standards for Differential Privacy. And this is also why the decision to abandon disclosure control based on realistic threat models in favor of Differential Privacy is irrational.<sup>17</sup>

---

<sup>17</sup> Indeed, the Census Bureau’s decision may not satisfy even constitutional rational basis review. Courts in recent years have found that regulations on the sale of caskets or on the practices of hair braiding studios had so little connection to societal welfare or risk mitigation that even these types of economic regulations imposed by statute were unconstitutional. *See St. Joseph Abbey v. Castille*, 712 F.3d 215 (5th Cir. 2013) (striking down a Louisiana law that gave funeral homes exclusive rights to sell caskets that the state attempted to justify by “abstraction for hypothesized ends”); *Brantley v. Kuntz*, 98 F. Supp. 3d 884 (W.D. Tex. 2015) (finding that regulations requiring salons to have sinks and certain types of equipment were irrational as applied to African hair braiding studio).

### **III. Neither Law Nor Public Distrust Can Justify the Census Bureau’s Decision to Adopt Differential Privacy**

Finally, there are no provisions in the U.S. Census Act that require the Census Bureau to take the action it has, nor are there any crises in public trust that can justify a dramatic shift in privacy protocols.

#### **A. Privacy Laws**

Privacy laws have long been crafted to allow data to be shared broadly or publicly in statistical, deidentified format despite the inherent risks involved. The “reasonableness” standard is the approach embodied in federal privacy regulations and industry guidance documents, and it is particularly well-matched to public data. HIPAA, for example, applies only to personal health information “(i) that identifies the individual; or (ii) [w]ith respect to which there is a reasonable basis to believe that the information can be used to identify the individual.” 45 C.F.R. §169.103. Subsequent guidance and regulations make clear that traditional disclosure avoidance techniques meet the standard as long as individuals cannot be re-identified under realistic assumptions of threat. DHS guidance documents on HIPAA compliance do not require or even recommend the use of Differential Privacy.<sup>18</sup> (If they had, management of the pandemic would have been particularly chaotic.)

The language in the Census Act is similar to HIPAA’s. The relevant confidentiality provision reads:

---

<sup>18</sup> DEPT. HEALTH & HUMAN SERVS., GUIDANCE REGARDING METHODS FOR DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION IN ACCORDANCE WITH THE HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT (HIPAA) PRIVACY RULE (2012), available at <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard>.

§ 9. Information as confidential; exception

(a) Neither the Secretary, nor any other officer or employee . . . may . . .

(2) make any publication whereby the data furnished by any particular establishment or individual under this title *can be identified* . . . .

13 U.S.C. § 9 (emphasis added). Although the act does not contain the phrase “there is a reasonable basis to believe...” the operative language is nearly identical to HIPAA. Both laws ask whether information can be used to identify individuals. The phrase “reasonable basis for belief” provides a mental state requirement (negligence), but even if the Census Act intends to impose a strict liability regime on Census Bureau officers and employees, the task of assessing which types of data can or cannot be identified is the same as the HIPAA context. Moreover, if this weren’t the case—if the phrase were meant to prohibit *any* publications that have any hypothetical chance of causing identification, Differential Privacy with any budget above 0 would violate the Act just as surely as traditional disclosure control techniques do.

It unlikely, however, Congress intended to impose a strict liability rule in any case. An excessively cautious approach to privacy would permit government agencies to evade public accountability and would close off the social benefits of public access. These are the core purposes of the Census Bureau.

**B. Public Trust**

The Census Bureau’s adoption of Differential Privacy could also make sense if a spate of successful reidentification attacks warranted a new approach to data privacy, but there is no such history, and there is no public outcry about the statistical data products routinely released by the Census Bureau. If anything, the use of Differential Privacy could spur public distrust and resentment by injecting doubt in the accuracy and reliability of data used to allocate resources and

define voting districts. The Bureau is legally obligated not only to protect the confidentiality of the Census records, but also to protect the vitality and accuracy of the information in its possession. Section 181 of the US Census Act requires the census to produce “current, comprehensive, and reliable data” for state, county, and local government purposes. 13 U.S.C. § 181.

The public has as much interest in the reasonable accuracy of statistical census data as in its reasonable privacy. Indeed, the Census Bureau promises respondents that “Federal funds, grants and support to states, counties and communities are based on population totals and breakdowns by sex, age, race and other factors. Your community benefits the most when the census counts everyone. When you respond to the census, you help your community gets its fair share of the more than \$675 billion per year in federal funds spent on schools, hospitals, roads, public works and other vital programs.”<sup>19</sup>

Differential Privacy undermines the Census Bureau’s mission of collecting and providing reliable and credible information, and as the public becomes aware of the significant damage done to the accuracy of data, a crisis in trust is likely to emerge. For example, the Census Bureau imposes several constraints on their use of Differential Privacy so that negative numbers of people are not reported. But the combination of non-negativity and the state-level population invariants consistently leads to bias in the reporting of counts for small subgroups.<sup>20</sup> To reduce the bias, at least one Census Bureau advisor has suggested the Bureau should consider dropping the non-negativity constraint even though “it may be confusing to say that a town has a negative, fractional

---

<sup>19</sup> U.S. Census Bureau, *Why a Census?: How the Census Benefits Your Community*, <https://www.census.gov/programs-surveys/decennial-census/2020-census/about/why.html>

<sup>20</sup> Barber Expert Report, Doc. 3-5 at 13-14 (explaining that “[t]he combination of the non-negativity constraint and population invariants consistently leads to bias increasing counts of small subgroups and small geographic units and decreasing counts of larger subgroups and geographic units.” (citation omitted))).



number of individuals with a particular combination of uncommon attributes”.<sup>21</sup> Negative numbers of people in the official statistics is more than confusing, though. Political fights are already suffering from a dearth of shared facts. By using Differential Privacy, the Census Bureau is putting one of the few sources of ground truth at risk. When Americans see Census Bureau reports with 141,000 Alabama children living without parents, *see* Declaration of Thomas Bryan, Doc. 3-6 at 11, distrust and low response rates are likely to ensue.

#### **IV. The Census Bureau’s Position Sets a Trap for Public Records Laws**

The U.S. Census Bureau’s claim that Differential Privacy is the *only* defensible way to keep statistical data safe sets a terrible precedent for government transparency and accountability more generally. The effect on public records laws could be devastating. If government agencies are able to justify their decisions to withhold records because they are not “differentially private” (even if they can be deidentified quite well), the landscape of public records and government accountability would change for the worse. These changes will offend American democratic values regardless of political identity. While public universities might use privacy exemptions to avoid public controversy related to Affirmative Action, law enforcement agencies will use those same exemptions to avoid public controversy related to racially biased policing.

For example, for several years the New York Civil Liberties Union suspected that the New York Police Department (NYPD) was using stop & frisk procedures in racially discriminatory ways. Aggregated reports had already verified that the number of police stops and frisks were growing, but the organization was not able to provide convincing evidence of racial bias until 2011, when the group successfully sued the NYPD under New York’s freedom of information law

---

<sup>21</sup> Michael B. Hawes, *U.S. Census Bureau, Implementing Differential Privacy: Seven Lessons From the 2020 United States Census*, HARV. DATA SCI. REV., Issue 2.2 (Apr. 30, 2020), <https://perma.cc/DB66-9B5R>.

to gain access to an individual-level database documenting the stop and frisk program. This data provided strong circumstantial evidence of intensive policing of minorities without reasonable suspicion and without any meaningful gains in safety, and it provided the impetus and basis for a civil rights challenge against the NYPD.<sup>22</sup> Yet the data contained in the NYPD database could have been used to reidentify a stopped individual using the location, timing, and demographic characteristics of the individuals who were stopped.<sup>23</sup>

Other public records litigation has required police departments to provide footage from body-worn cameras (with faces blurred) and has required U.S. Customs and Border Patrol to provide redacted documents about individuals stopped at internal checkpoint or by roving patrol in order to facilitate citizen and journalist investigations of potential abuses.<sup>24</sup> Public records disclosures of individual-level data has allowed journalists to find flaws in state sex abuse criminal cases<sup>25</sup> and evidence that children from low-income households were excluded from public gifted

---

<sup>22</sup> *Stop-And-Frisk 2011*, NEW YORK CIV. LIBERTIES UN. (May 2012), [https://www.nyclu.org/sites/default/files/publications/NYCLU\\_2011\\_Stop-and-Frisk\\_Report.pdf](https://www.nyclu.org/sites/default/files/publications/NYCLU_2011_Stop-and-Frisk_Report.pdf); *Stop-and-Frisk in the de Blasio era*, NEW YORK CIV. LIBERTIES UN. (Mar. 2019) [https://www.nyclu.org/sites/default/files/field\\_documents/20190314\\_nyclu\\_stopfrisk\\_singles.pdf](https://www.nyclu.org/sites/default/files/field_documents/20190314_nyclu_stopfrisk_singles.pdf); *Floyd v. City of New York*, 959 F. Supp. 2d 540 (S.D. N.Y. 2013).

<sup>23</sup> Footage of several NYPD stops are available on YouTube, some with date and location information.

<sup>24</sup> Jessie Gomez, *Louisiana judge grants access to state policy body-camera footage*, MUCKROCK (Mar. 1, 2019) <https://www.muckrock.com/news/archives/2019/mar/01/louisiana-bodycam/>; James Lyall, et al., *Record of Abuse, Lawlessness and Impunity in Border Patrol's Interior Enforcement Operations*, AM. CIV. LIBERTIES UN. OF ARIZ., 4 (Oct. 2015); *ACLU Found. of Ariz. v. U.S. Dep't Homeland Sec.*, No. CV-14-02052-TUC-RM (BPV), 2017 WL 8895339 (D. AZ. Jan. 26, 2017) (rejecting the government's reidentification risk argument).

<sup>25</sup> Tennessee Watson, *Justice Isn't Always Done for Child Sex Abuse-I Know Firsthand*, REVEAL (Aug. 11, 2016), <https://revealnews.org/article/tennessee-watson-justice-isnt-always-done-for-child-sexual-abuse-i-know-firsthand/>.

and talented education programs.<sup>26</sup> Public access to data of this sort will be under grave threat if state agencies are able to say, as the Census Bureau has, that accuracy must be compromised for the sake of an abstract and baffling concept of privacy.<sup>27</sup>

## CONCLUSION

All statistical data carries risk of inadvertent disclosure. Those who prepare public data must find a sensible way to balance the risks of privacy invasion against the risks of not allowing public access and accountability. The disclosure avoidance literature routinely acknowledges that data de-identification is a balancing act between data privacy and research utility, and it has served the U.S. Census very well up to this point.<sup>28</sup> Differential Privacy introduces unreasonable amounts

---

<sup>26</sup> Joseph Neff, Ann Doss Helms, & David Raynor, *Why Have Thousands of Smart, Low-Income NC Students Been Excluded from Advanced Classes?*, THE CHARLOTTE OBSERVER (May 21, 2017), <https://www.charlotteobserver.com/news/local/education/article150488822.html>.

<sup>27</sup> Government agencies have already used privacy as an excuse to withhold domestic violence data and now-infamous photographs from Abu Ghraib. Caitlin Russell, *In Massachusetts, laws intended to protect domestic abuse victims' privacy are being used to deny access to data about enforcement*, MUCKROCK (Jan. 9, 2018), <https://www.muckrock.com/news/archives/2018/jan/09/dv-mass-data/>; *ACLU v. Dep't of Defense*, 543 F.3d 59, 84 (2d Cir. 2008) (“According to the defendants, when combined with information contained in the investigative reports associated with the detainee images, release of the photographs could make it possible to identify the detainees.”)

<sup>28</sup> “Stewards of social data [] face a fundamental tension. At one extreme, a data steward could share a complete dataset publicly with everyone. This *full release* approach maximizes the potential for scientific discovery, but it also maximizes risk to the people whose information is in the dataset. At the other extreme, a data steward could share the data with no one. This *no release* approach minimizes risk to participants, but it also eliminates benefits that could come from the responsible use of the data. In between these two extremes—no release and full release—there are a variety of intermediate solutions, which involve balancing risk to participants and benefits to science.” Ian Lundberg, et al., *Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge* (Sept. 10, 2019), <https://journals.sagepub.com/doi/10.1177/2378023118813023>; See also Mark Elliot & Josep Domingo-Ferrer, *The Future of Statistical Disclosure Control*, NAT’L STATISTICIAN’S QUALITY REV. (2018) (“SDC fundamentally consists of two processes: disclosure risk analysis and disclosure control. Controlling the disclosure risk must be done in a way that optimizes the trade-off between risk and utility. While risk must be kept below the maximum acceptable threshold (set by law or by good practices), utility must be kept above the minimum threshold that data users can accept. Without utility constraints, there would be no reason to control disclosure: one might rather

of inaccuracy while making allowances for privacy loss that are not based on real world risk. Thus, the switch from grounded risk-based privacy precautions to the abstract guarantees provided by Differential Privacy is an arbitrary and capricious abuse of the Census Bureau’s discretion. Although the Census Bureau has significant freedom to exercise its judgment over how best to balance privacy and data utility, the agency still “must examine the relevant data and articulate a satisfactory explanation for its action including a rational connection between the facts found and the choice made.” *Motor Vehicle Mfrs. Ass’n of U.S., Inc. v. State Farm Mut. Auto. Ins. Co.*, 463 U.S. 29, 30 (1983). Because Differential Privacy is, by design, insensitive to real world probabilities and risks, and because there is no history of significant privacy breaches that would be corrected by a change to Differential Privacy, the Census Bureau cannot meet even the generous standard that applies to agency discretionary judgments.

Respectfully submitted,

/s/ Christopher W. Weller  
CHRISTOPHER W. WELLER (WEL020)

*Counsel for Amicus Curiae*  
*Professor Jane Bambauer*

**OF COUNSEL:**

CAPELL & HOWARD, P.C.  
150 South Perry Street  
Montgomery, AL 3104  
Phone: (334) 241-8066  
Fax: (334) 241-8266  
chris.weller@chlaw.com

---

suppress the data entirely, which would result in 0% disclosure risk!”); El Emam & Luk Arbuckle, ANONYMIZING HEALTH DATA: CASE STUDIES AND METHODS TO GET YOU STARTED 28 (2013) (“Zero risk can’t guarantee if we want to share any useful data. The very small risk is the trade-off we need to accept to realize the many important benefits of sharing and using health data... Regulators don’t expect zero risk either—they accept that a very small risk is reasonable.”)

**CERTIFICATE OF SERVICE**

I hereby certify that on 9<sup>th</sup> day of April, 2021, I filed with the Court and served on all counsel through the CM/ECF system the foregoing document.

STEVE MARSHALL  
*Attorney General of Alabama*  
Edmund G. LaCour Jr. (ASB-9182-U81L)  
*Solicitor General*  
A. Barrett Bowdre (ASB-2087-K29V)  
*Deputy Solicitor General*  
James W. Davis (ASB-4063-I58J)  
Winfield J. Sinclair (ASB-1750-S81W)  
Brenton M. Smith (ASB-1656-X27G)  
*Assistant Attorneys General*

STATE OF ALABAMA  
OFFICE OF THE ATTORNEY GENERAL  
501 Washington Ave.  
Montgomery, AL 36130  
Telephone: (334) 242-7300  
Fax: (334) 353-8400  
Edmund.LaCour@AlabamaAG.gov  
Barrett.Bowdre@AlabamaAG.gov  
Jim.Davis@AlabamaAG.gov  
Winfield.Sinclair@AlabamaAG.gov  
Brenton.Smith@AlabamaAG.gov  
*Counsel for the State of Alabama*

Jason B. Torchinsky (VA Bar No. 47481)\*  
Jonathan P. Lienhard (VA Bar No. 41648)\*  
Shawn T. Sheehy (VA Bar No. 82630)\*  
Phillip M. Gordon (VA Bar. No. 95621)\*  
HOLTZMAN VOGEL JOSEFIAK  
TORCHINSKY, PLLC  
15405 John Marshall Hwy  
Haymarket, VA 20169  
(540) 341-8808 (Phone)  
(540) 341-8809 (Fax)  
Jtorchinsky@hvjt.law  
Jlienhard@hvjt.law  
Ssheehy@hvjt.law  
Pgordon@hvjt.law  
*\*pro hac vice*  
*Counsel for Plaintiffs*

BRIAN M. BOYNTON  
Acting Assistant Attorney General  
ALEXANDER K. HAAS  
Director, Federal Programs Branch  
BRAD P. ROSENBERG  
Assistant Director, Federal Programs Branch  
ZACHARY A. AVALLONE  
ELLIOTT M. DAVIS (N.Y. Reg. No. 4596755)  
JOHN ROBINSON  
Trial Attorneys  
Civil Division, Federal Programs Branch  
U.S. Department of Justice  
1100 L St. NW  
Washington, DC 20005  
Phone: (202) 514-4336  
Fax: (202) 616-8470  
E-mail: [elliott.m.davis@usdoj.gov](mailto:elliott.m.davis@usdoj.gov)  
*Counsel for Defendants*

/s/ Christopher W. Weller  
Of Counsel

Case No. 3:21-cv-211-RAH-ECM-KCN

---

---

**IN THE UNITED STATES DISTRICT COURT  
FOR THE MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION**

---

**THE STATE OF ALABAMA, *et al.*,**

Plaintiffs,

v.

**UNITED STATES DEPARTMENT OF COMMERCE, *et al.*,**

Defendants.

---

**AMICUS CURIAE BRIEF IN SUPPORT OF PLAINTIFFS FROM  
(1) SENATE OF PENNSYLVANIA REPUBLICAN CAUCUS,  
(2) PENNSYLVANIA SENATE PRESIDENT PRO TEMPORE JAKE  
CORMAN, AND  
(3) PENNSYLVANIA SENATE MAJORITY LEADER KIM WARD**

---

Albert L. Jordan  
Jonathan A. Griffith  
Wallace, Jordan, Ratliff, & Brandt, L.L.C.  
800 Shades Creek Parkway, Suite 400  
Birmingham, Alabama 35209  
Phone: (205) 870-0555  
Fax: (205) 871-7534  
bjordan@wallacejordan.com  
jgriffith@wallacejordan.com

*Attorneys for Senate of Pennsylvania Republican Caucus,  
Pennsylvania Senate President Pro Tempore Jake Corman,  
and Pennsylvania Senate Majority Leader Kim Ward*

---

---

**TABLE OF CONTENTS**

TABLE OF AUTHORITIES ..... ii

STATEMENT OF INTEREST OF AMICI CURIAE.....1

ARGUMENT .....5

    I. The traditional tabulation of census data, unlawfully delayed and altered in the name of differential privacy, is needed for fair apportionment of representation, as well as reliable planning of local funding streams.....6

        A. The Census Bureau’s Demonstration Files reveal differential privacy’s use generates inaccurate and unreliable population statistics. .... 7

        B. Erroneous census data would impair or eliminate the equitable distribution of funds received under Pennsylvania’s grant, loan, and funding programs and create distrust of state and local government..... 9

    II. The deadline for delivery of intrastate census population data set out in § 141(c) means what it says, and is due to be enforced as written.....13

CONCLUSION .....14

CERTIFICATE OF SERVICE .....15

**TABLE OF AUTHORITIES**

<b><u>Cases</u></b>	<b><u>Page(s)</u></b>
<i>Agre v. Wolf</i> , 284 F. Supp. 3d 591) (E.D. Pa. 2018).....	4
<i>Albert v. 2001 Legislative Reapportionment Comm’n</i> , 790 A.2d 989 (Pa. 2002).....	5
<i>Commonwealth ex rel. Spencer v. Levin</i> , 293 A.2d 15 (Pa. 1972).....	4
<i>Corman v. Torres</i> , 287 F. Supp. 3d 558 (M.D. Pa. 2018).....	4
<i>Donatelli v. Casey</i> , 826 F. Supp. 131 (E.D. Pa. 1993).....	4
<i>Erfer v. Commonwealth</i> , 794 A.2d 325 (Pa. 2002).....	4
<i>Garcia v. 2011 Legislative Reapportionment Comm’n</i> , 559 Fed. Appx. 128 (3d Cir. 2014).....	5
<i>Garcia v. 2011 Legislative Reapportionment Comm’n</i> , 938 F. Supp. 2d 542 (E.D. Pa. 2013).....	5
<i>Harrison v. Pennsylvania Legislative Reapportionment Com.</i> , 1992 U.S. Dist. LEXIS 5315 (E.D. Pa. April 21, 1992) .....	7
<i>Holt v. 2011 Legislative Reapportionment Comm’n</i> , 38 A.3d 711 (Pa. Jan. 2012).....	5
<i>Holt v. 2011 Legislative Reapportionment Comm’n</i> , 67 A.3d 1211 (Pa. 2013).....	5
<i>In re 1991 Pennsylvania Legislative Reapportionment Com.</i> , 609 A.2d 132 (Pa. 1992).....	4
<i>In re Pennsylvania Congressional Dist. Reapportionment Cases</i> , 535 F. Supp. 191 (M.D. Pa. 1982).....	4



*In re Pennsylvania Congressional Dist. Reapportionment Cases*,  
567 F. Supp. 1507 (M.D. Pa. 1982).....4

*In re Reapportionment Plan for the Pennsylvania General Assembly*,  
442 A.2d 661 (Pa. 1982).....4

*League of Women Voters of Pa. v. Commonwealth*,  
175 A.3d 282 (Pa. Jan. 22, 2018) .....4

*League of Women Voters of Pa. v. Commonwealth*,  
178 A.3d 737 (Pa. Feb. 7, 2018).....4, 9

*League of Women Voters of Pa. v. Commonwealth*,  
181 A.3d 1083 (Pa. Feb. 19, 2018).....4, 9

*Mellow v. Mitchell*,  
607 A.2d 204 (Pa. 1992).....4

*Pileggi v. Aichele*,  
843 F. Supp. 2d 584 (E.D. Pa. 2012).....5

*Precision Mktg., Inc. v. Com., Republican Caucus of the Sen. of PA*,  
78 A.3d 667 (Pa. Cmwlt. 2013).....2

*Vieth v. Pa.*,  
188 F. Supp. 2d 532 (M.D. Pa. 2002).....4

*Vieth v. Pa.*,  
195 F. Supp. 2d 672 (M.D. Pa. 2002).....4

*Vieth v. Pa.*,  
241 F. Supp. 2d 478 (M.D. Pa. 2003).....4

**Statutes, Rules, and Constitutional Provisions** **Page(s)**

2 U.S.C. § 2c .....3

13 U.S.C. § 141(c) .....6, 13

13 U.S.C. § 141 .....1

Pa. Const. Art. I, § 1.....2  
Pa. Const. Art. II, § 9 .....2  
Pa. Const. Art. II, § 16 .....1  
Pa. Const. Art. II, § 17(b).....2  
Pa. Const., Art. II, § 17(a).....3  
Pa. Const., Art. II, § 17(c).....3  
Pa. Const., Art. IV, § 4.....2  
  
12 Pa.C.S. § 3401.....12  
64 Pa.C.S. § 1557(e)(2).....11  
64 Pa.C.S §1551.....12  
35 P.S. § 751.10 .....11  
72 P.S. § 400.2508 .....11  
72 P.S. § 1602-D .....11  
72 P.S. § 2615.7(b).....10  
72 P.S. § 8802-C .....11  
72 P.S. § 8822-G.....11  
  
Public Law 94-171 .....7  
  
Rule 5, Rules of the Senate of Pennsylvania .....7

*Amici Curiae*, the Senate of Pennsylvania Republican Caucus (the “Caucus”), Senate Majority Leader Kim Ward, and Senate President Pro Tempore Jake Corman file this brief in support of Plaintiffs, the State of Alabama, Robert Aderholt, in his official and individual capacities; William Green; and Camaran Williams.

**STATEMENT OF INTEREST OF AMICI CURIAE**

The three amici curiae submitting this brief are interested in this action by virtue of duties imposed by their official positions in the government of the State of Pennsylvania, and the duties imposed by the State’s Constitution. In brief, their official positions call on them to rely on data provided by the U.S. Department of Commerce through its Bureau of the Census for performing their duties. As shown in more detail below, census data is used for figuring proper intrastate reapportionment of legislative districts as well as proper distribution of certain public funds.

All three amici are, in the words of the Census Act, “officers or public bodies having responsibility for legislative reapportionment or districting of [the] State.” 13 U.S.C. § 141. The Caucus is composed of all Republican members of the Senate of the State of Pennsylvania. Under Article II, § 16 of the Pennsylvania Constitution, the Senate is composed of 50 members. The Senate is one part of the General Assembly of Pennsylvania in which “the legislative power of this

Commonwealth” is vested, under Article II, § 1 of the State Constitution. The Caucus was created with the Senate’s constitutional authority under Article II of the State Constitution. At present, the Caucus is composed of 27 Senate members, and one Independent who caucuses with Republicans. The Caucus is said to be “an integral constituent of the Senate” and to perform “essential legislative functions and administrative business in the Senate.” *Precision Mktg., Inc. v. Com., Republican Caucus of the Sen. of PA*, 78 A.3d 667, 675 (Pa. Cmwlth. 2013).

The President Pro Tempore Jake Corman (“PPT”) is an officer of the State Senate, as established by Art. II, § 9 of the State Constitution. Subject to election by the full Senate, the PPT serves as the President of the Senate in the absence of the Lieutenant Governor. Pa. Const., Art. IV, § 4. The PPT is also responsible, along with the Speaker of the House, for certifying the four (4) legislative members of the Legislative Reapportionment Commission under Pennsylvania’s Constitution. Pa. Const. Art. II, § 17(b).

The Majority Leader Kim Ward is elected by vote of the Caucus. According to the Rules of the Senate, the Majority Leader serves as President of the Senate in the absence of the Lieutenant Governor and of the PPT. *See* Rules of the Senate of Pennsylvania, Rule 5 (adopted Jan. 5, 2021). In addition to her role with the Senate, the Majority Leader is a member of the Commonwealth’s Legislative Reapportionment Commission. Pa. Const. Art. II, § 17(b).

The State Senate’s lawmaking power, and therefore part of the official duties of members of the Caucus, includes the establishment of district lines for the members of Congress elected from Pennsylvania. *See* 2 U.S.C. § 2c (“there shall be established by law a number of districts equal to the number of Representatives to which such State is so entitled . . . .”). In addition, the boundaries of the districts, from which Senators are elected, are adjusted by a Legislative Reapportionment Commission “in each year following the year of the Federal decennial census.” Pa. Const., Art. II, § 17(a), (c).

The Plaintiffs’ complaint and memorandum of law submitted in support of their motion, as well as other amici, have detailed a great number of harms that will result from the Census Bureau’s use of differential privacy. Without this Court’s intervention, identical and similar harms will occur across the nation, including in the Commonwealth of Pennsylvania.

Amici share the concerns the Plaintiffs have detailed in their Complaint, and thus wishes to inform the Court of certain other injuries that the use of differential privacy will inflict. In short, the inaccurate and delayed census data will significantly harm communities’ planning capabilities, funding streams, and political environments, and needlessly generate a substantial amount of litigation centered on state legislative redistricting.

The litigation experience of Pennsylvania during the last several decades over census-dependent redistricting shows the depth of the interest about the issues raised here. The effect extends to establishing districts for both Congressional seats<sup>1</sup>, as well as the districts from which the members of its Senate (as well as the members of its House) are elected.<sup>2</sup> That experience shows a continuing interest in

---

<sup>1</sup> See *In re Pennsylvania Congressional Dist. Reapportionment Cases*, 535 F. Supp. 191 and 567 F. Supp. 1507 (M.D. Pa. 1982) (refusing to preliminarily and later permanently enjoin the congressional redistricting plan following the 1980 decennial census); *Mellow v. Mitchell*, 607 A.2d 204 (Pa. 1992) (choosing from among six plans submitted by various elected officials because the General Assembly had not timely passed legislation approving a map); *Donatelli v. Casey*, 826 F. Supp. 131 (E.D. Pa. 1993)(holding that the temporary representation of a district by an individual no longer residing in the district as a result of redistricting did not violate the state or federal Constitutions, pending the expiration of the official’s term of office); *Erfer v. Commonwealth*, 794 A.2d 325 (Pa. 2002) (upholding Act 1, the General Assembly’s legislation redrawing congressional districts following the 2000 Census); *Vieth v. Pa.*, 188 F. Supp. 2d 532 and 195 F. Supp. 2d 672 (M.D. Pa. 2002) (declaring 2000 congressional redistricting plan unconstitutional and ordering the General Assembly to prepare a revised plan); *Vieth v. Pa.*, 241 F. Supp. 2d 478 (M.D. Pa. 2003) (upholding the GA’s supplemental redistricting plan passed following the Court’s Order in *Vieth I*); *League of Women Voters of Pa. v. Commonwealth*, 175 A.3d 282 (Pa. Jan. 22, 2018) (striking down the Congressional Redistricting Act of 2011 and ordering delivery of a new plan by the General Assembly no later than Feb. 9, 2018—approximately 18 days from the date of the Court’s order); *League of Women Voters of Pa. v. Commonwealth*, 178 A.3d 737 (Pa. Feb. 7, 2018) (opinion in support of Jan 22 Order); *League of Women Voters of Pa. v. Commonwealth*, 181 A.3d 1083 (Pa. Feb. 19, 2018) (adopting a reapportionment plan for federal Congressional districts generated by the Court in light of the General Assembly’s “failure” to “timely” submit a revised plan following the Court’s January 22, 2018 Order); *Agre v. Wolf*, 284 F. Supp. 3d 591 (E.D. Pa. 2018) (court rejected claims of partisan gerrymandering on grounds it’s a non-justiciable political question); *Corman v. Torres*, 287 F. Supp. 3d 558 (M.D. Pa. 2018) (federal court rejected request by legislators and elected officials to enjoin the use of the Court’s redistricting plan following League of Women Voters).

<sup>2</sup> See *Commonwealth ex rel. Spencer v. Levin*, 293 A.2d 15 (Pa. 1972) (encompassing seventeen (17) cases challenging the plan, but upholding the Commission’s final plan); *In re Reapportionment Plan for the Pennsylvania General Assembly*, 442 A.2d 661 (Pa. 1982) (encompassing twenty-nine (29) cases challenging the plan, but upholding the Commission’s final plan); *In re 1991 Pennsylvania Legislative Reapportionment Com.*, 609 A.2d 132 (Pa. 1992) (encompassing twenty-five (25) cases challenging the plan, but upholding the

the timely and accurate receipt of census data—as Alabama and the other Plaintiffs seek here.

For these reasons, the Caucus has the kind of interest that the Court grant the Plaintiffs’ motion to preliminarily enjoin the Defendants’ from both implementing differential privacy and delaying the provision of accurate census data. Anything less will result in deliberately late and false population tabulations, which would not only be useless, but would actually inflict great damage to our State’s political and financial wellbeing.

### **ARGUMENT**

These amici urge the Court to grant the declaratory and injunctive relief, or alternatively, mandamus relief, as sought by the Plaintiffs. The Defendants have “specific tabulations of population” due to be “reported to the Governor . . . and to the officers or public bodies having responsibility for legislative apportionment or

---

Commission’s final plan); *Harrison v. Pennsylvania Legislative Reapportionment Com.*, 1992 U.S. Dist. LEXIS 5315 (E.D. Pa. April 21, 1992) (court rejected challenge to final redistricting plan); *Albert v. 2001 Legislative Reapportionment Comm’n*, 790 A.2d 989 (Pa. 2002) (encompassing eleven (11) cases challenging the plan, but upholding the Commission’s final plan); *Pileggi v. Aichele*, 843 F. Supp. 2d 584 (E.D. Pa. 2012) (use of the 2001 plan adopted by the LRC was appropriate pending final resolution of the post-2010 decennial census reapportionment plan); *Holt v. 2011 Legislative Reapportionment Comm’n*, 38 A.3d 711 (Pa. Jan. 2012) and 38 A.3d 711 (Pa. Feb 2012) (Holt I) (encompassing twelve (12) cases challenging the plan, and remanding the plan to the Legislative Reapportionment Commission on a finding that the final plan was contrary to law); *Holt v. 2011 Legislative Reapportionment Comm’n*, 67 A.3d 1211 (Pa. 2013) (Holt II) (upholding the plan created by the Commission following the 2012 remand in Holt I) *Garcia v. 2011 Legislative Reapportionment Comm’n*, 938 F. Supp. 2d 542 (E.D. Pa. 2013) (rejecting challenges to the 2011 final plan and application to the 2013 and 2014 election cycles) and *Garcia v. 2011 Legislative Reapportionment Comm’n*, 559 Fed. Appx. 128 (3d Cir. 2014) (rejecting challenge to redistricting plan on standing grounds).

districting.” 13 U.S.C. § 141(c). Moreover, the deadline for these “tabulations of population” is “within one year after the decennial census date.” (*Id.*). That date passed on March 31. Nonetheless, Defendants in the name of enhancing privacy of census respondents are alleged to have abandoned the administrative techniques used in 2000 and 2010. Instead, they are arranging with a technique known as “differential privacy” to provide false data for the intrastate “tabulations” required by § 141(c). *See* Complaint at 19–31 (Mar. 10, 2021) (Doc. 1). For the reasons set out below, these amici object to this change of course, and request the Court not allow it.

**I. The traditional tabulation of census data, unlawfully delayed and altered in the name of differential privacy, is needed for fair apportionment of representation, as well as reliable planning of local funding streams.**

In 2019, the Pennsylvania State Data Center (“PaSDC”) submitted a report (attached as “Exhibit A”) to the United States Census Bureau outlining many of the harmful effects the Census Bureau’s use of differential privacy would cause within the Commonwealth of Pennsylvania. The PaSDC (established in 1981 by executive order of the governor of Pennsylvania) serves as the state’s official source of population and economic statistics, and as the state’s liaison to the Census Bureau. The PaSDC serves businesses, non-profits, government agencies, and individuals, answering more than 15,000 requests for information each year. It also assists the



Pennsylvania Legislative Reapportionment Commission and the General Assembly with census data analysis for districting purposes.

The PaSDC's 2019 report followed a 2019 Census Bureau preview of how the differential privacy algorithm would distort 2020 census data by applying the algorithm to 2010 census data. That application resulted in what are known as the "Demonstration Files," consisting of the demonstration version of Public Law 94-171 (which requires the Census Bureau to provide states opportunity to identify the small area geography for which they need data for legislative redistricting) and selected tables from the proposed 2020 Demographic and Housing Characteristics Summary File ("Summary File") for all states, Puerto Rico, and the District of Columbia. The PaSDC then compared the Summary File data for Pennsylvania's counties and county subdivisions (i.e., municipalities), and compared the Public Law 94-171 redistricting data for Pennsylvania's state legislative districts. Below, the Caucus outlines the PaSDC's findings, which are as alarming as they informative, and demonstrates how inaccurate census data derails Pennsylvania's ability to accurately and fairly operate its state loans, grants, and funding programs.

**A. The Census Bureau's Demonstration Files reveal differential privacy's use generates inaccurate and unreliable population statistics.**

The PaSDC's comparison of the Demonstration Files with original 2010 census data revealed that differential privacy causes considerable deviations in

population tabulations at the county subdivision level.<sup>3</sup> In fact, the differential privacy algorithm caused the populations of at least 84 county subdivisions to reflect an increase of more than twenty percent from the original 2010 census data. The algorithm doubled (or more than doubled) the population of 12 of the communities. Conversely, at least 37 county subdivisions lost over 20% of their populations. *See* (Ex. A at 2).

Differential privacy inflated persons-per-household statistics, distorted age cohorts (five-year ranges in age, such as 35–39 or 40–44) to show zero members of more than half of the age cohorts in 175 county subdivisions and zero members in 25% of the age cohorts in another 730 communities. *See (Id. at 2–3)*. The PaSDC also noted many county subdivisions experienced significant differences in their racial makeup—those where the single race alone represented two percent or more of the total population higher as a result of differential privacy than in the original 2010 census data. *See (Id.)*.

Additionally, the PaSDC found that differential privacy changed the total population numbers in most of Pennsylvania’s state Senate and House districts. For State House Districts, 98 lost population and 105 gained population, with decreases as high as 655 persons lost and increases as high as 771 persons gained. For State

---

<sup>3</sup> A graph depicting the differences for all municipalities is available at <https://public.tableau.com/views/DifferentialPrivacyandMunicipalPopulations/DFandPAMunicipalities>

Senate Districts, 28 lost population while 22 gained. The largest population decrease was 815 persons while the largest increase was 1,321 persons. *See (Id.* at 3). As noted in explaining the interest of these amici, if past is prologue, Pennsylvania's redistricting process is sure to be highly scrutinized in the courts again, as evident by the recent court ruling, *League of Women Voters v. Commonwealth of Pennsylvania*, 178 A. 3d 737 and 181 A.3d 1083(Pa. 2018), implementing remedial congressional districts. The distorted numbers, if allowed, will only inject erroneous data and confusion into that scrutiny, and significantly increase tensions and the likelihood of litigation.

**B. Erroneous census data would impair or eliminate the equitable distribution of funds received under Pennsylvania's grant, loan, and funding programs and create distrust of state and local government.**

Many—if not all—of Pennsylvania's communities depend on an accurate reporting of Decennial Census data because such data is criteria for eligibility for several of Pennsylvania's grant, loan, and funding programs. Therefore, distorted and inaccurate population tabulations would significantly affect whether and how much funding certain communities could receive. The PaSDC provided many compelling examples of how distorted census data would directly harm many of Pennsylvania's county subdivisions, especially those with limited or declining resources.

***The Municipal Liquid Fuels Program***

The Municipal Liquid Fuels Program (“MLF”) makes funding available to local governments (i.e., county subdivisions) to support construction, reconstruction, maintenance, and repair of public roads or streets. Because Pennsylvania relies on accurate census data to determine how Liquid Fuel funds are to be distributed to the state’s more than 2,500 communities, the program’s method of distributing funds is one of the most visible examples of how differential privacy would directly (and arbitrarily) affect our State’s local governments and communities. *See* 72 P.S. § 2615.7(b) (Requiring population calculations for apportionment to be based on most recent census figures).

MLF funds are no small issue. Local governments across Pennsylvania profoundly depend on them for a variety of activities in their local area, not the least of which is maintaining their roads in the harsh Pennsylvania climate, where freeze/thaw cycles are destructive. Applying the differential privacy algorithm to the 2010 census data reduces the total population of 1,200 local governments, resulting in a redistribution of \$2.4 million of MLF funds (based on false numbers rather than actual population). To the state’s smaller communities, especially those that are economically depressed, even small losses of such funds would certainly be harmful. That demonstrable redistribution directly thwarts Pennsylvania’s

ability to accurately and fairly operate its programs and distribute state tax dollars in a logical and equitable fashion.

***Other state grant, loan, and funding programs***

Inaccurate 2020 census population data would also frustrate Pennsylvania's ability to operate many other grant and loan programs. For example, the City Revitalization Improvement Zone ("CRIZ") provides opportunities to spur new growth, helps revive downtowns, and creates jobs for local residents. The program develops "pilot zones" based on areas of a certain geographic size and a population of at least 7,000. *See* 72 P.S. § 8802-C. Of course, the false data resulting from the use of differential privacy jeopardizes (and makes it impossible to predict) the eligibility of CRIZ benefits for municipalities in that population range.

There is no question that differential privacy's inaccurate population totals would severely impact county subdivisions' eligibility for and receipt of various programs, loans, and grants. The only question is which subdivisions will suffer as a result of the incorrect data, and to what degree.<sup>4</sup>

---

<sup>4</sup> *See, e.g.*, 72 P.S. § 1602-D (Codifying the Local Government Capital Project Loan Program, which denies eligibility for its low-interest loans to local governments whose population exceed \$12,000); 64 Pa.C.S. § 1557(e)(2) (PA Venture Capital Investment Program, requiring at least 50% of program's funding be spent in areas with populations of 1,000,000 or less); 73 P.S. § 400.2508 (Community Development Bank Loan Program - eligibility based on whether county population declined by at least 10 percent outside of metropolitan areas); 35 P.S. § 751.10 (Infrastructure Investment Authority - limiting funding for improvements to lesser of \$1,000 per resident or \$10,000,000); 72 P.S. § 8822-G (Rural Jobs and Investment Tax Credit Program - eligibility based on population thresholds of 50,000).

Similarly, some programs within Pennsylvania consider population decline to prioritize eligibility—which, of course, cannot be accurately determined based on inaccurate population numbers.<sup>5</sup> For example, PaSDC found that “one hundred communities in Pennsylvania that would have reported a population increase in 2010 under the original [2010 census data] would report a decrease in 2010 under the Demonstration Files (using differential privacy).” (Ex. A at 6). “[O]ver two hundred communities that would have reported a population decrease in 2010 under the original [2010 census data] would report an increase in 2010” using differential privacy. (*Id.*). Therefore, differential privacy would thwart the state’s desire to prioritize those communities most in need because the algorithm would cause their populations to be falsely inflated to the point of appearing to have experienced a population increase.

To conclude, the use of differential privacy needlessly distorts crucial population data at the county subdivision level. Consequently, because Pennsylvania statutes require state grant, loan, and funding programs to distribute tax dollars according to accurate census data, the use of differential privacy precludes the government’s ability to comply. The use of differential privacy will

---

<sup>5</sup>*See, e.g.*, 12 Pa.C.S. § 3401 (Infrastructure and Facilities Improvement Program, providing grants to issuers of debt to assist with payment of debt service—guidelines available at: <https://dced.pa.gov/programs/infrastructure-and-facilities-improvement-program-ifip/>); 64 Pa.C.S §1551 (Business in Our Sites Grants and Loans, helping communities attract growing businesses—guidelines available at <https://dced.pa.gov/programs/business-in-our-sites-grants-and-loans-bos/>).

exchange Pennsylvania's logical and purposeful distribution of grants, loans, and other funds to its taxpayers for a new distribution based on arbitrary and distorted census data. Not only will such a method thwart the policies and strategies behind such distribution, it will also sow distrust in the minds of the state's taxpayers regarding how their tax dollars are being apportioned.

**II. The deadline for delivery of intrastate census population data set out in § 141(c) means what it says, and is due to be enforced as written.**

The delay in the release of population data until September in derogation of the March 31 deadline imposed by 13 U.S.C. § 141(c) will squeeze unduly the Pennsylvania General Assembly in its work. There will be only five short months—until mid-February 2022 when candidates begin circulating petitions for ballot listing—to create a Congressional district plan, enact a statute adopting the plan, and potentially litigate it. At least, that tightened schedule will operate unless there is a legislative change in the 2022 primary election. Likewise the Legislative Apportionment Commission will face similar problems in establishing a plan for the Senate and House districts of the General Assembly. Again, in light of the extensive litigation history within Pennsylvania, a compact time frame invites an increase in litigation that places significant pressure on the legislature, Redistricting Commission, Courts, and candidates. For these reasons, these amici urge the Court to conclude that § 141(c) means what it says, and should not be allowed, in effect, to be re-written by administrative officials.

**CONCLUSION**

For these reasons, these amici request the Court grant relief to the Plaintiffs, as requested in their Complaint.

Respectfully submitted on April 12, 2021.

*s/ Albert L. Jordan*

---

Albert L. Jordan

bjordan@wallacejordan.com

*s/ Jonathan A. Griffith*

---

Jonathan A. Griffith

jgriffith@wallacejordan.com

**OF COUNSEL:**

Wallace, Jordan, Ratliff & Brandt, L.L.C.

800 Shades Creek Parkway, Suite 400

Birmingham, Alabama 35208

Phone: (205) 870-0555

Fax: (205) 871-7534

*Attorneys for Senate of Pennsylvania Republican Caucus,  
Pennsylvania Senate President Pro Tempore Jake Corman,  
and Pennsylvania Senate Majority Leader Kim Ward*



**CERTIFICATE OF SERVICE**

I certify that on April 12, 2021, I electronically filed the foregoing with the Clerk of Court using the CM/ECF system, which will send notification of such filing to:

Brian M. Boynton  
*Acting Assistant Attorney General*  
Alexander K. Haas  
*Director, Federal Programs Branch*  
Brad P. Rosenberg  
*Assistant Director, Federal Programs Branch*  
Elliott M. Davis  
Zachary A. Avallone  
John Robinson  
*Trial Attorneys*  
Civil Division, Federal Programs Branch  
U.S. Department of Justice  
1100 L St. NW  
Washington, DC 20005  
elliott.m.davis@usdoj.gov

Jason B. Torchinsky  
Jonathan P. Lienhard  
Shawn T. Sheehy  
Phillip M. Gordon  
Holtzman Vogel Josefiak Torchinsky,  
PLLC  
15405 John Marshall Hwy  
Haymarket, VA 20169  
jtorchinsky@hvjt.law  
jlienhard@hvjt.law  
ssheehy@hvjt.law  
pgordon@hvjt.law

Steve Marshall  
*Attorney General of Alabama*  
Edmund G. LaCour Jr.  
*Solicitor General*  
A. Barrett Bowdre  
*Deputy Solicitor General*  
James W. Davis  
Winfield J. Sinclair  
Brenton M. Smith  
*Assistant Attorneys General*  
State of Alabama  
Office of the Attorney General  
501 Washington Avenue  
Montgomery, Alabama 36130  
Edmund.LaCour@AlabamaAG.gov  
Barrett.Bowdre@AlabamaAG.gov  
Jim.Davis@AlabamaAG.gov  
Winfield.Sinclair@AlabamaAG.gov  
Brenton.Smith@AlabamaAG.gov

Christopher W. Weller  
Capell & Howard, P.C.  
150 South Perry Street  
Montgomery, AL 36104  
Chris.weller@chlaw.com

Rik S. Tozzi  
Ryan J. Hebson  
Attorneys for Amici States  
BURR & FORMAN LLP  
420 North 20th Street, Suite 3400  
Birmingham, Alabama 35203  
rtozzi@burr.com  
rhebson@burr.com

SEAN D. REYES  
*Attorney General of Utah*  
MELISSA A. HOLYOAK\*  
*Solicitor General*  
STATE OF UTAH  
OFFICE OF THE ATTORNEY  
GENERAL  
350 N. State Street, Suite 230  
P.O. Box 142320  
Salt Lake City, UT 84114-2320  
melissaholyoak@agutah.gov  
\*pro hac vice application forthcoming

*s/ Albert L. Jordan*  
\_\_\_\_\_  
OF COUNSEL

# Exhibit A

# **Differential Privacy and the 2020 Census in PA: Analyses and Concerns**

The Pennsylvania State Data Center

Jennifer Shultz, [jjb131@psu.edu](mailto:jjb131@psu.edu)

Tim Schock, [trs69@psu.edu](mailto:trs69@psu.edu)

## ANALYSIS OF DEMONSTRATION DATA

The Pennsylvania State Data Center (PaSDC) has compared the Demonstration files released for 2010 utilizing the Disclosure Avoidance System (DAS, or “Differential Privacy”) to the original 2010 data for both the Summary File and P.L. 1994 Redistricting Data. The Summary File data were compared for Pennsylvania’s counties and county subdivisions (i.e., municipalities) while the P.L. 1994 Redistricting Data were compared for Pennsylvania’s state legislative districts.

### County and Subcounty Analysis

We found no significant differences at the county level of analysis. County subdivisions, however, showed considerable differences. We have visualized these differences for all municipalities in Pennsylvania, which can be viewed at:

<https://public.tableau.com/views/DifferentialPrivacyandMunicipalPopulations/DFandPAMunicipalities>

Overall, we found that at least 84 county subdivisions had populations that increased by more than 20 percent from the original Summary File to the Demonstration product. Additionally, population increased so drastically in 12 communities that their populations doubled (or more than doubled). Conversely, at least 37 county subdivisions lost over 20% of their populations. This trend, combined with differences in published housing unit counts, also inflated persons per household statistics which ranged from one person-per-household to seven persons-per-household higher in the DAS file when compared to the original Summary File.

Age cohorts were severely distorted under in the DAS data. In over 175 county subdivisions, more than half of the age cohorts (five-year ranges) for men and women were “zeroed out” – we use this phrase to explain the phenomenon we observed where five-year age cohorts that had data in the original Summary File had zero population in the DAS file (see Figure 1). These areas, which were the most impacted, all shared small populations of 1,000 or less.

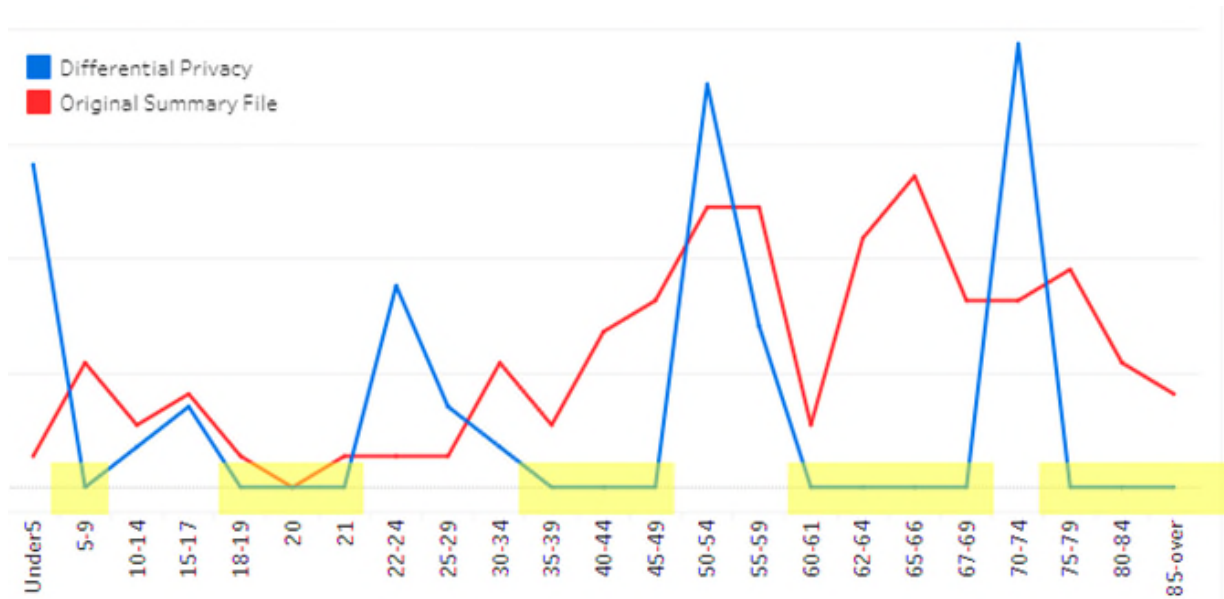


Figure 1. Example of “zeroed out” cohorts for the male population of Grove Township, Cameron County, PA.

The “zeroing” of population cohorts occurred in 25 percent of the age cohorts for an additional 730 communities. This phenomenon was not observed in over 300 county subdivisions, all having a population of 2,500 or more. The effect of DAS on small communities is alarming, considering that most of Pennsylvania’s county subdivisions have small populations.

We also compared the racial distributions of communities according to the DAS Demonstration file and the original Summary File. Table 1 highlights the number of county subdivisions that experienced significant differences in their racial fabric (here, we define significant differences as those where the single race alone represented two percent or more of the total population higher in the DAS file than it did in the original Summary File), and the nature of those differences (whether there were more or less individuals of that race/ethnicity).

**Table 1.** Comparison of racial distribution of communities in DAS Demonstration data to original Summary File data.

Race/Ethnicity*	Significant Difference	Increased	Decrease
American Indian/Native Alaskan	14	10	4
Asian	32	28	4
Black or African American	96	61	35
Native Hawaiian/Other Pacific Islander	-	-	-
Multiracial	163	128	35
White	368	109	259
Hispanic or Latino	163	115	48

\*Single races represent non-Latino individuals. Hispanic or Latino individuals are of any race.

**State Legislative District Analysis**

We found differences even for total population for both Senate and House districts in Pennsylvania. We have visualized these differences for all Senate and House districts in Pennsylvania, which can be viewed at:

<https://public.tableau.com/views/DifferentialPrivacyandRedistricting/Comparison>

We found that total population changed in most districts. For State House Districts, 98 lost population and 105 gained population. The largest population decrease was as high as 655 persons lost while the largest population increase was as high as 771 persons gained. For State Senate Districts, 28 lost population while 22 gained. The largest population decrease was as high as 815 persons lost while the largest population increase was as high as 1,321 persons gained.

The 2020 redistricting process in Pennsylvania is sure to be highly scrutinized, as is evident by the recent court ruling that implemented remedial state legislative districts. The Disclosure Avoidance System will add additional tensions to the process as participants must defend the quality of the data.

## **IMPACT ON COMMUNITY FUNDING IN PENNSYLVANIA**

Several of Pennsylvania's grant, loan, and funding programs related to community and economic development activities across the state use Decennial Census data as criteria for eligibility. Changes in the reported total population of communities in Pennsylvania may affect the amount of funding they receive from a program or exclude their eligibility altogether.

### **The Municipal Liquid Fuels Program**

Perhaps the best example of funding that is derived from Decennial counts is the Municipal Liquid Fuels Program (MLF). The program makes funding available to local governments (i.e., county subdivisions) to support construction, reconstruction, maintenance, and repair of public roads or streets. Liquid Fuel funds are distributed to communities across the state (of which there are over 2,500) based on each community's total population and the total mileage of roads.

Local governments across the state deeply depend on these moneys for a variety of activities in their local area, not the least of which is maintaining their roads in a Pennsylvania climate, where freeze/thaw cycles destroy roads. This program is one of the most visible ways in which the Census can impact a community.

Each person, or unit of population, was valued just under \$20 in the most recent allocation of Liquid Fuels funds. With the implementation of the Disclosure Avoidance System, just under 1,200 local governments have lower populations than originally reported in 2010. The most significant loss occurs for Philadelphia, whose loss of 1,500 individuals equates to a loss of over \$30,000. In total, over \$2.4 million dollars would need to be redistributed among local governments in Pennsylvania under the Disclosure Avoidance System.

### **Total Population as Thresholds**

Many programs use total population, based on the most recent Decennial Census, as a threshold to determine eligibility. Communities' participation in the following programs could be affected by any changes to their population:

- *City Revitalization Improvement Zone (CRIZ)* – Provides opportunities to spur new growth, helping to revive downtowns and create jobs for the residents in the regions. Vacant, desolate, underutilized or abandoned space will be developed, thereby creating jobs, increasing personal incomes, growing state and local tax revenues, reviving local economies and improving the lives of city residents and visitors. (<https://dced.pa.gov/programs/city-revitalization-improvement-zone-criz/>)
  - This program develops "pilot zones" based on areas of a certain geographic size and a population of at least 7,000. In Pennsylvania, this would have disqualified Clairton city, Allegheny County, Pennsylvania whose population under the original Summary File data was 7,021 and decreased to 6,986 in the Demonstration data. Many of Pennsylvania's communities have populations near 7,000.
- *Local Government Capital Project Loan Program (LGCP)* – Provides low-interest loans to local governments for equipment and facility needs. (<https://dced.pa.gov/programs/local-government-capital-project-loan-program-lgcpl/>)

- This program defines eligibility based on a community having a population size of 12,000 or less. Many of Pennsylvania's communities have populations near 12,000.
- *PA Venture Capital Investment Program* – Provides loans to venture capital partnerships to invest in growth-stage PA companies. (<https://dced.pa.gov/programs/new-pa-venture-capital-investment-program/>)
  - This program specifies that at least 50 percent of the programs total funding should be spent in area of Pennsylvania outside the Philadelphia MSA and with populations of 1,000,000 or less.
- *Community Development Bank Loan Program* – Provides debt financing for Community Development Financial Institutions (CDFIs). (<https://dced.pa.gov/programs/pennsylvania-community-development-bank-loan-program/>)
  - This program defines eligibility based on whether service areas have a total combined population that exceeds its metropolitan area, and also whether a county population has declined by at least 10 percent outside of metropolitan areas.
- *Infrastructure Investment Authority (PennVEST)* – Provides low-interest loans and grants for new construction or for improvements to publicly or privately-owned drinking water, storm water or sewage treatment facilities, as well as non-point source pollution prevention best management practices. PENNVEST also provides loan funding to remediate brownfields sites, as well as loan funding to individual homeowners for repair or replacement of their malfunctioning on-lot septic system or first-time connection to a public sewer collection system. The Advance Funding Program provides low-interest loans to provide funding for the design and engineering needed to improve water and wastewater management systems. (<https://dced.pa.gov/programs/pennsylvania-infrastructure-investment-authority-pennvest/>)
  - This program provides funding for communities that should not exceed \$1,000 per resident of the community or \$10,000,000, whichever is less.
- *Rural Jobs and Investment Tax Credit Program (RITC)* – Offers rural business owners access to capital for business development in rural areas. The capital is sourced to Rural Growth Funds, designated to receive up to \$100 million dollars in capital contributions from investors. The Commonwealth of Pennsylvania is using this investment tool to attract and retain rural businesses to the commonwealth, create family sustaining jobs, and to stimulate economic growth in rural businesses. (<https://dced.pa.gov/programs/rural-jobs-and-investment-tax-credit-program-rjtc/>)
  - This program defines eligibility based on areas of the state that is not in a city whose population of 50,000 or more or an urbanized area adjacent to a city that has a population of 50,000 or more.

Slight changes to the reported population of communities across Pennsylvania, specifically those not due to real population growth or decline but instead due to the Census's Disclosure Avoidance System, could wholly exclude them from program funding, as shown by the programs and thresholds above.

### **Identifying Declining Populations**

Additionally, several programs use population decline to determine eligibility. Specifically, the language of these programs' guidelines dictate that the community "... is located in an area with a particular for economic development, as shown by ... declining population...".

- *Business in Our Sites Grants and Loans (BOS)* – Empowers communities to attract growing and expanding businesses by helping them build an inventory of ready sites. (<https://dced.pa.gov/programs/business-in-our-sites-grants-and-loans-bos/>)



- *Infrastructure and Facilities Improvement Program (IFIP)* – A multi-year grant program that will provide grants to certain issuers of debt in order to assist with the payment of debt service. (<https://dced.pa.gov/programs/infrastructure-and-facilities-improvement-program-ifip/>)
- *Pipeline Investment Program (PIPE)* – Provides grants to construct the last few miles of natural gas distribution lines to business parks, existing manufacturing and industrial enterprises, which will result in the creation of new economic base jobs in the commonwealth while providing access to natural gas for residents. (<https://dced.pa.gov/programs/pipeline-investment-program/>)
- *Tax Increment Financing (TIF) Guarantee Program* – Promotes and stimulates the general economic welfare of various regions and communities in the Commonwealth and assists in the development, redevelopment and revitalization of Brownfield and Greenfield sites in accordance with the TIF Act. The program provides credit enhancement for TIF projects to improve market access and lower capital costs through the use of guarantees to issuers of bonds or other indebtedness. (<https://dced.pa.gov/programs/tax-increment-financing-tif-guarantee-program/>)

As such, we fear the degree to which the implementation of the Disclosure Avoidance System will impact consistency in comparisons across Decennial counts. For example, when comparing the percentage change in communities (i.e., county subdivisions) from 2000 to 2010, over 300 municipalities had percentage changes using the Differential Privacy Demonstration Data that were opposite of the direction of change using the original 2010 Summary File data.

In other words, one hundred communities in Pennsylvania that would have reported a population increase in 2010 under the original Summary File data would report a decrease in 2010 under the Differential Privacy Demonstration data, and over two hundred communities that would have reported a population decrease in 2010 under the original Summary File data would report an increase in 2010 under the Differential Privacy Demonstration data.

### **Library funding**

The Pennsylvania State Data Center is currently working with the Pennsylvania Department of Education to provide demographic data for library service areas of public libraries in Pennsylvania. Library service areas are defined by combining county subdivisions or parts of county subdivisions. This data is used in developing programming, planning outreach and maintaining collections. The severe distortion of age cohorts noted above would be especially impactful in this analysis. Libraries are planning programming for children, seniors and other specific age cohorts based on this data.

### **CONCERNS**

Total population is a fundamental baseline for Pennsylvania's communities related to planning, funding, and political representation. The census tract may be the standard geographic level for the Bureau, but how DAS will affect real-world geographies must be considered, especially for a Commonwealth such as Pennsylvania.

As we have shown, the DAS data inflate, deflate, and in some cases reverse the extent of the population change in county subdivisions across the state of Pennsylvania. These changes distort communities' planning capabilities, funding streams, and political environments. We are concerned

that these changes could diminish the trust of communities across the state as it relates to the accuracy and reliability of Census data.

Of specific importance is the distortion to age and race distributions seen in the demonstration data products at the county subdivision and legislative district levels. We expect that this same distortion would be seen in other off-spine geographies such as school districts. As noted, this data is critical for state and local government funding and planning.

We appreciate this opportunity to analyze the demonstration products and provide feedback. Please contact us if you have any questions.

**UNITED STATES DISTRICT COURT FOR THE  
MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION**

THE STATE OF ALABAMA; ROBERT ADERHOLT, Representative for Alabama’s 4th Congressional District, in his official and individual capacities; WILLIAM GREEN; and CAMARAN WILLIAMS,

Plaintiffs,

v.

UNITED STATES DEPARTMENT OF COMMERCE; GINA RAIMONDO, in her official capacity as Secretary of Commerce; UNITED STATES BUREAU OF THE CENSUS, an agency within the United States Department of Commerce; and RON JARMIN, in his official capacity as Acting Director of the U.S. Census Bureau,

Defendants.

CIVIL ACTION NO.  
3:21-cv-211-RAH-ECM-KCN

BRIEF OF *AMICI CURIAE* STATE OF UTAH AND 15 OTHER STATES IN SUPPORT OF PLAINTIFFS

**INTRODUCTION**

The States of Utah, Alaska, Arkansas, Florida, Kentucky, Louisiana, Maine, Mississippi, Montana, Nebraska, New Mexico, Ohio, Oklahoma, South Carolina, Texas, and West Virginia (*Amici States*) agree with Plaintiffs that the Secretary’s intended use of differential privacy deprives states of accurate “[t]abulations of population” of state subparts to use in legislative apportionment and districting under 13 U.S.C. § 141(c). *Amici States* also agree that the Secretary can comply with the privacy requirements of 13 U.S.C. § 9 by alternative methods that do not deprive the states of the numbers to which section 141 entitles them. They submit this *amicus* brief to explain the detrimental effects that using the differential privacy method would have on both re-districting and administering state and federal programs.

## ARGUMENT

### **I. Utah’s analysis of the 2010 demonstration data shows that differential privacy will result in inaccurate 2020 subpopulation data affecting redistricting and state and federal program funding.**

In October 2019, the Census Bureau released demonstration data to permit states to review the effects of differential privacy. *See* 2010 Demonstration Data Products, <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html>. The demonstration data included the census data from 2010 that was treated with the new differential privacy method. *Id.* Using a mathematical model, the Census Bureau injects “noise”—false information—into the raw data to minimize the risk of privacy disclosure. *Id.* The Utah State Legislature analyzed the 2010 demonstration data, comparing it with the previously received 2010 redistricting data and sent its findings to the Census Bureau. *See* Letter of the Utah State Legislature (Feb. 13, 2020), [https://www.ncsl.org/Portals/1/Documents/Redistricting/UT\\_Differential\\_Privacy\\_%28Signed%29.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/UT_Differential_Privacy_%28Signed%29.pdf).

The Utah State Legislature identified three major harms from using differential privacy for census data. *Id.* at 1. *First*, it would make accurate redistricting at the local level impossible. The analysis showed that when differential privacy was applied to the 2010 data, there was a statewide net loss of nearly 15,000 people from Utah’s cities and towns, including two cities that lost 50% of their populations. *Id.* Indeed, with inaccurate subpopulation data, the State would be unable to accurately receive and distribute funds to localities. Like many states, Utah has state revenue-sharing statutes and receives federal funding based on population formulas derived from census data. Inaccurate data would “impact state and federal funding that is disbursed in compliance with” those statutes and formulas. *Id.* at 1.

*Second*, inaccurate data could “adversely affect longitudinal studies about health, safety and welfare.” *Id.* If the academic and professional policy analyses that legislators rely on to inform public policy decisions were based on inaccurate data, the Legislature could no longer rely on them, and would have to essentially legislate in the dark. *Id.* And *third*, because of the population shifts, the Utah Legislature expressed concerns that the State would not be able to fulfill its constitutional obligation to satisfy population and equality requirements in redistricting. *Id.* at 2.

These concerns remained even after the Census Bureau tweaked the data. The Bureau released additional sets of demonstration data in May 2020, September 2020, and November 2020, modifying the amount of injected “noise” with each dataset. *See* <https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx>. The Utah Legislature analyzed the November 2020 data in the same way it had analyzed the modified 2010 data. *See* Differential Privacy, Utah State Legislature (March 2021) (“2021 Utah Report”) (attached at Exh. 1). While it saw an improvement from the October 2019 to the November 2020 demonstration data, it did not cure the population inaccuracies. For example, Congressional districts three and four had populations increase and decrease by nearly 50 voters, respectively, *id.* at 32—significantly higher than the one-person-one-vote principle requiring states to draw legislative districts that are nearly equivalent in population. *See Evenwel v. Abbott*, 136 S. Ct. 1120, 1123-24 (2016).

As with its analysis of the modified 2010 data, the Utah Legislature’s concerns with the November 2020 data went beyond redistricting. The Utah Legislature observed that while the November 2020 data improved, there remained “some significant population changes, particularly in rural municipalities.” 2021 Utah Report, Exh. 1 at 1. Specifically, several cities suffered a population decrease of over 30%. *Id.* at 13. And inaccurate data would translate to lost funding for those

communities. For example, in FY2017, Utah received \$9 billion from census-guided federal funding. *See* Andrew Reamer, *Counting for Dollars 2020: The Role of the Decennial Census in the Geographic Distribution of Federal Funds*, Brief 7: Comprehensive Accounting of Census-Guided Federal Spending (FY2017): Part B: State Estimates at 3, <https://perma.cc/MUP5-6KJ5>. Nationwide, \$1.5 trillion was distributed through 316 federal spending programs on 2010 census-derived data. *Id.* at 1. Thus, like Utah, inaccurate subpopulation data will harm distribution of census-guided funding in all states.

**II. Other states’ analyses also recognize the harm differential privacy will inflict on rural areas and minority racial groups.**

Utah is not alone in its concerns about redistricting, funding, and data accuracy. All states use census data to redistrict, obtain and distribute federal funds, and administer many state and local programs. Because differential privacy creates false information—by design—it prevents the states from accessing municipal-level information crucial to performing this essential government functions. And the distorting impact of differential privacy will likely fall hardest on some of the most vulnerable populations—rural areas and minority racial groups. *See* National Conference of State Legislatures, *Differential Privacy for Census Data Explained*, Mar. 15, 2021, available at <https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx>.

As one University of Virginia researcher explained in a letter to Governor Northam, skewing minority group data is particularly problematic when a State must accommodate majority-minority districts. *See* Memorandum from Meredith Strohm Gunter to Hon. Ralph Northam, Jan. 23, 2020, available at [https://www.ncsl.org/Portals/1/Documents/Redistricting/VA\\_CensusDistortionProgram\\_VAGovernor\\_2020-01-23.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/VA_CensusDistortionProgram_VAGovernor_2020-01-23.pdf). Because the “noise-injected proxy” would change the “actual size of the voting age population in each census block” as well as its racial

characteristics, “[m]ajority-minority districts could lose their status” or a non-minority-majority district might mistakenly have majority-minority status conferred upon it. *Id.*

California’s leaders recently sent a letter to the White House Chief of Staff expressing concerns that inaccuracies introduced by differential privacy would “hamper the ability of states and localities to establish political districts that comply with the United States Constitution’s ‘one-person, one-vote’ principle and with the protections of the Voting Rights Act of 1965.” Feb. 2021 Letter from California leaders to Ronald Klain, available at [https://www.ncsl.org/Portals/1/Documents/Redistricting/California\\_Differential\\_Privacy\\_summary2021.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/California_Differential_Privacy_summary2021.pdf). A joint analysis from Asian Americans Advancing Justice and Mexican American Legal Defense and Educational Fund explained that this would likely lead to minorities being underrepresented. *See* Preliminary Report: Impact of Differential Privacy & the 2020 Census on Latinos, Asian Americans, and Redistricting, available at <https://advancingjustice-aaajc.org/report/preliminary-report-impact-differential-privacy-2020-census-latinos-asian-americans>.

Other states also shared concerns about funding equity for localities and data accuracy. As the Virginia researcher explained, myriad state programs—from housing and transportation to emergency management—rely on accurate data to deliver state services to those who need it. [https://www.ncsl.org/Portals/1/Documents/Redistricting/VA\\_CensusDistortionProgram\\_VA-Governor\\_2020-01-23.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/VA_CensusDistortionProgram_VA-Governor_2020-01-23.pdf). And legislators rely on census-derived statistics to calibrate programs for those in need. *Id.*

Two officials from Maine—its state economist and data center lead—expressed similar concerns in a letter to the Census Bureau’s director, explaining that their analysis showed that “small, rural places suffer the most” from inaccurate estimates.” Feb. 20, 2020 Letter to Steven

Dillingham, available at [https://www.ncsl.org/Portals/1/Documents/Redistricting/ME\\_Letter\\_to\\_Census\\_on\\_differential\\_privacy\\_concerns\\_Maine\\_SDC.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/ME_Letter_to_Census_on_differential_privacy_concerns_Maine_SDC.pdf). Washington State’s state demographer wrote a similar letter to the Bureau’s director about the outsized impact that rural areas would suffer under differential privacy, saying that the data would be “unusable for large parts of” the state and skew funding away from small towns. Feb. 6, 2020 Letter to Steven Dillingham, [https://www.ncsl.org/Portals/1/Documents/Redistricting/WA\\_OFM\\_DAS\\_Response\\_Letter.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/WA_OFM_DAS_Response_Letter.pdf). He found the error rate “alarmingly high” and “extremely problematic” for state functions. *Id.* The Colorado General Assembly echoed similar redistricting, funding, and data accuracy concerns to those of other states—though in their analysis, the data skewed in favor of rural areas. *See* June 1, 2020 Letter to Steven Dillingham, available at [https://www.ncsl.org/Portals/1/Documents/Elections/CO\\_State\\_Legislative\\_Leadership\\_Letter.pdf?ver=2020-08-04-132435-780&timestamp=1596569177678](https://www.ncsl.org/Portals/1/Documents/Elections/CO_State_Legislative_Leadership_Letter.pdf?ver=2020-08-04-132435-780&timestamp=1596569177678).

Finally, demographic researchers from the University of California Riverside and the University of Washington did four case studies using data from Alaska to illustrate just how strange the local-level results of using differential privacy can be. They found that three population blocks included several children and no adults; 1,252 voting blocks switched from having one or more persons of voting age to having no persons of voting age; 830 blocks went the other way, from having no persons of voting age to having at least one; and that 96% of blocks (12,366 of 12,870) with one or more inhabitants showed a different number of persons. Population Association of America, *The Effect of Differential Privacy Disclosure Avoidance System Proposed by the Census Bureau on 2020 Census Products: Four Case Studies of Census Blocks in Alaska*, available at <https://www.populationassociation.org/blogs/paa-web1/2021/03/30/the-effect-of-the-differential-privacy-disclosure>.



*Amici* States share concerns that the Bureau's proposed use of differential privacy will harm State redistricting, funding, and data collection. This in turn will harm all the States' citizens, but the burden will fall disproportionately on minorities and rural areas. This Court should rule in favor of the Plaintiffs.

Dated: April 13, 2021

Respectfully submitted,

/s/ Ryan J. Hebson

Rik S. Tozzi (TOZ001)  
ASB-7144-Z48R  
Ryan J. Hebson (HEB003)  
ASB-3200-R74H  
*Attorneys for Amici States*  
BURR & FORMAN LLP  
420 North 20th Street, Suite 3400  
Birmingham, Alabama 35203  
Telephone: (205) 251-3000  
Facsimile: (205) 458-5100  
rtozzi@burr.com  
rhebson@burr.com

SEAN D. REYES  
*Attorney General of Utah*  
MELISSA A. HOLYOAK\* (Utah Bar No. 9832)  
*Solicitor General*  
STATE OF UTAH  
OFFICE OF THE ATTORNEY GENERAL  
350 N. State Street, Suite 230  
P.O. Box 142320  
Salt Lake City, UT 84114-2320  
Telephone: (801) 538-9600  
melissaholyoak@agutah.gov  
\**pro hac vice* application pending

TREG R. TAYLOR  
ALASKA ATTORNEY GENERAL

LESLIE RUTLEDGE  
ARKANSAS ATTORNEY GENERAL

ASHLEY MOODY  
FLORIDA ATTORNEY GENERAL

DANIEL CAMERON  
KENTUCKY ATTORNEY GENERAL

JEFF LANDRY  
LOUISIANA ATTORNEY GENERAL

AARON M. FREY  
MAINE ATTORNEY GENERAL

LYNN FITCH  
MISSISSIPPI ATTORNEY GENERAL

AUSTIN KNUDSEN  
MONTANA ATTORNEY GENERAL

DOUGLAS J. PETERSON  
NEBRASKA ATTORNEY GENERAL

HECTOR BALDERAS  
NEW MEXICO ATTORNEY GENERAL

DAVE YOST  
OHIO ATTORNEY GENERAL

MIKE HUNTER  
OKLAHOMA ATTORNEY GENERAL

ALAN WILSON  
SOUTH CAROLINA ATTORNEY GENERAL

KEN PAXTON  
TEXAS ATTORNEY GENERAL

PATRICK MORRISEY  
WEST VIRGINIA ATTORNEY GENERAL

**CERTIFICATE OF SERVICE**

I hereby certify that I have served a copy of the foregoing document by Notice of Electronic Filing, or, if the party served does not participate in Notice of Electronic Filing, by U.S. First Class Mail on this the 13th day of April, 2021:

STEVE MARSHALL  
*Attorney General of Alabama*  
Edmund G. LaCour Jr.  
*Solicitor General*  
A. Barrett Bowdre  
*Deputy Solicitor General*  
James W. Davis  
Winfield J. Sinclair  
Brenton M. Smith  
*Assistant Attorneys General*

STATE OF ALABAMA  
OFFICE OF THE ATTORNEY GENERAL  
501 Washington Ave.  
Montgomery, Alabama 36104  
Telephone: (334) 242-7300  
Facsimile: (334) 353-8400  
edmund.lacour@alabamaag.gov  
barrett.bowdre@alabamaag.gov  
jim.davis@alabamaag.gov  
winfield.sinclair@alabamaag.gov  
brenton.smith@alabamaag.gov

Christopher W. Weller  
CAPELL & HOWARD, P.C.  
150 South Perry Street  
Montgomery, Alabama 36104  
Telephone: (334) 241-8000  
Facsimile: (334) 241-8266  
chris.weller@chlaw.com

Jason B. Torchinsky  
Jonathan P. Lienhard  
Shawn T. Sheehy  
Phillip M. Gordon  
HOLTZMANVOGEL JOSEFIAK  
TORCHINSKY PLLC  
15405 John Marshall Hwy  
Haymarket, Virginia 20169  
Telephone: (540) 341-8808  
Facsimile: (540) 341-8809  
jtorchinsky@hvjt.law  
jlienhard@hvjt.law  
ssheehy@hvjt.law  
pgordon@hvjt.law

BRIAN M. BOYNTON  
*Acting Assistant Attorney General*  
ALEXANDER K. HAAS  
*Director, Federal Programs Branch*  
BRAD P. ROSENBERG  
*Assistant Director, Federal Programs Branch*  
Zachary A. Avallone  
Elliott M. Davis  
John Robinson  
*Trial Attorneys*  
United States Department of Justice  
Civil Division, Federal Programs Branch  
1100 L Street, N.W.  
Washington, DC 20005  
Telephone: (202) 616-8489  
zachary.a.avallone@usdoj.gov  
elliott.m.davis@usdoj.gov  
john.j.robinson@usdoj.gov

/s/ Ryan J. Hebson  
OF COUNSEL

# **Exhibit 1**

***[Differential Privacy, Utah State  
Legislature (March 2021)]***

**Differential Privacy** is a term used by the U.S. Census Bureau to describe a privacy technique that scrambles census data at the census block level in order to protect the personally identifiable information of census respondents. Although the Census Bureau has used privacy techniques since 1970, it has never used a privacy technique that alters data as much as differential privacy.

### **U.S. Census Bureau**

In addition to conducting a complete and accurate enumeration of the United States every 10 years, the Census Bureau is also required by federal law to keep all personally identifiable information collected during the census, such as age, race, gender, marital status, etc., confidential.

### **Privacy Techniques Cracked**

Using other public data sets, complex algorithms, and super computers, it is possible for big data miners to reconstruct personally identifiable information from the census data.

### **New Privacy Strategy**

During the 2020 census, the Census Bureau intends to implement differential privacy for the first time. The Census Bureau reports that this technique is mathematically proven to protect personally identifiable information.

### **Privacy v. Accuracy**

The more privacy the Census Bureau protects, the less accurate the enumeration becomes. Less accurate data creates three concerns:

- State and local redistricting will be based on incorrect census block data;
- Distribution of federal and state monies may not reflect the actual population of the recipient municipalities;
- Academics, professional researchers, and policy analysts will make future policy recommendations regarding the health, safety, and welfare of individuals and the economy on inaccurate information.

### **Demonstration Data: October 2019 Version and November 2020 Version**

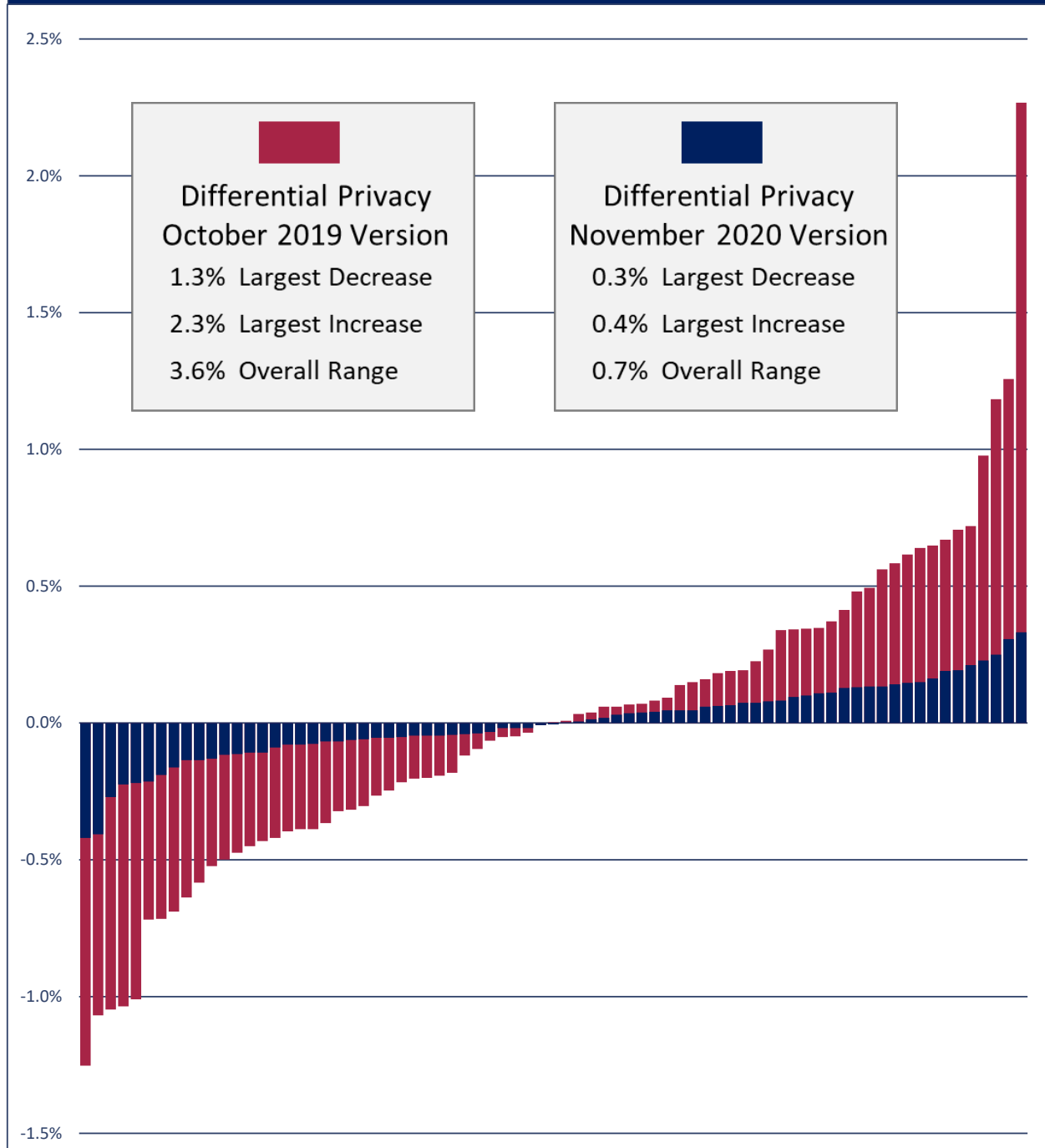
In this report, we refer to the Census Bureau's October 29, 2019 release of demonstration data as the "October Version" and the November 16, 2020 release as the "November Version." This report compares the differences that these two versions of differential privacy have on the populations within House, Senate, and Congressional districts and within the state's counties and municipalities, using the 2010 census data. The following maps and tables demonstrate how these two versions would have affected the 2010 census data.

### **Conclusion**

Although the information in this report does not predict variances in future census data, it does demonstrate how differential privacy techniques would have changed the 2010 census data. We conclude that although the November 2020 version represents a significant improvement over the October 2019 version, we still notice some significant population changes, particularly in rural municipalities. We also note that all population variances noted in this report are in addition to the unknown variances that were applied to the data by the Census Bureau.



### Differential Privacy Applied to 2010 House District Populations





## House Districts

### Differential Privacy Applied to 2010 House District Populations

#### Key to Colors



October 2019 Version				November 2020 Version			
District	2010 Redistricting Population	Number	Percent	District	2010 Redistricting Population	Number	Percent
11	36,871	-462	-1.25%	16	36,850	-155	-0.42%
29	36,853	-394	-1.07%	48	36,842	-150	-0.41%
43	36,857	-386	-1.05%	29	36,853	-100	-0.27%
74	36,874	-382	-1.04%	18	36,852	-83	-0.23%
21	36,832	-372	-1.01%	75	36,860	-81	-0.22%
40	36,855	-265	-0.72%	5	36,876	-79	-0.21%
6	36,851	-264	-0.72%	43	36,857	-70	-0.19%
56	36,852	-254	-0.69%	71	36,859	-60	-0.16%
30	36,858	-235	-0.64%	64	36,846	-50	-0.14%
64	36,846	-215	-0.58%	65	36,848	-50	-0.14%
13	36,859	-193	-0.52%	24	36,852	-48	-0.13%
44	36,847	-184	-0.50%	52	36,841	-43	-0.12%
67	36,859	-175	-0.47%	51	36,853	-42	-0.11%
14	36,873	-166	-0.45%	40	36,855	-40	-0.11%
34	36,851	-159	-0.43%	25	36,856	-40	-0.11%
35	36,860	-155	-0.42%	10	36,870	-33	-0.09%
45	36,856	-146	-0.40%	36	36,843	-29	-0.08%
51	36,853	-143	-0.39%	31	36,852	-29	-0.08%
46	36,854	-143	-0.39%	21	36,832	-28	-0.08%
52	36,841	-135	-0.37%	53	36,832	-25	-0.07%
25	36,856	-119	-0.32%	41	36,844	-25	-0.07%
60	36,851	-117	-0.32%	47	36,851	-23	-0.06%
22	36,862	-112	-0.30%	58	36,836	-22	-0.06%
10	36,870	-98	-0.27%	34	36,851	-20	-0.05%
62	36,839	-91	-0.25%	20	36,855	-20	-0.05%
48	36,842	-80	-0.22%	38	36,847	-19	-0.05%
42	36,857	-75	-0.20%	55	36,833	-17	-0.05%
32	36,839	-74	-0.20%	54	36,837	-17	-0.05%
31	36,852	-71	-0.19%	39	36,859	-17	-0.05%
41	36,844	-67	-0.18%	44	36,847	-16	-0.04%
54	36,837	-44	-0.12%	6	36,851	-15	-0.04%
20	36,855	-35	-0.09%	30	36,858	-14	-0.04%



## House Districts

7	36,855	-24	-0.07%	45	36,856	-12	-0.03%
8	36,867	-19	-0.05%	56	36,852	-7	-0.02%
2	36,847	-18	-0.05%	13	36,859	-7	-0.02%
16	36,850	-13	-0.04%	74	36,874	-7	-0.02%
61	36,853	-3	-0.01%	27	36,857	-3	-0.01%
72	36,846	1	0.00%	2	36,847	-2	-0.01%
36	36,843	3	0.01%	70	36,830	1	0.00%
59	36,844	12	0.03%	61	36,853	2	0.01%
19	36,874	14	0.04%	15	36,852	5	0.01%
17	36,871	22	0.06%	23	36,855	7	0.02%
33	36,845	22	0.06%	7	36,855	11	0.03%
66	36,857	25	0.07%	68	36,830	13	0.04%
63	36,855	26	0.07%	67	36,859	14	0.04%
4	36,844	30	0.08%	37	36,841	15	0.04%
47	36,851	34	0.09%	8	36,867	17	0.05%
3	36,852	51	0.14%	28	36,864	17	0.05%
5	36,876	55	0.15%	22	36,862	17	0.05%
23	36,855	59	0.16%	3	36,852	22	0.06%
15	36,852	67	0.18%	35	36,860	23	0.06%
38	36,847	70	0.19%	19	36,874	24	0.07%
18	36,852	71	0.19%	72	36,846	27	0.07%
65	36,848	83	0.23%	9	36,845	27	0.07%
1	36,851	99	0.27%	12	36,876	29	0.08%
27	36,857	125	0.34%	49	36,856	30	0.08%
69	36,830	126	0.34%	46	36,854	35	0.09%
75	36,860	127	0.34%	73	36,836	37	0.10%
70	36,830	128	0.35%	14	36,873	40	0.11%
57	36,854	137	0.37%	57	36,854	41	0.11%
58	36,836	152	0.41%	63	36,855	47	0.13%
55	36,833	177	0.48%	69	36,830	48	0.13%
39	36,859	182	0.49%	1	36,851	49	0.13%
28	36,864	207	0.56%	50	36,844	49	0.13%
71	36,859	215	0.58%	66	36,857	52	0.14%
24	36,852	227	0.62%	42	36,857	54	0.15%
12	36,876	236	0.64%	4	36,844	55	0.15%
9	36,845	239	0.65%	60	36,851	60	0.16%
49	36,856	247	0.67%	33	36,845	70	0.19%
53	36,832	260	0.71%	32	36,839	71	0.19%
50	36,844	265	0.72%	17	36,871	78	0.21%
37	36,841	360	0.98%	59	36,844	84	0.23%
26	36,850	436	1.18%	26	36,850	92	0.25%
73	36,836	463	1.26%	11	36,871	113	0.31%
68	36,830	835	2.27%	62	36,839	122	0.33%





## House Districts

### Differential Privacy Applied to 2010 House District Populations

#### Key to Colors

<span style="color: blue;">■</span> Less than -1.00%	<span style="color: lightcoral;">■</span> 0% to 0.40%
<span style="color: lightblue;">■</span> -1.00% to -0.40%	<span style="color: coral;">■</span> 0.40% to 1.00%
<span style="color: lightblue;">■</span> -0.40% to 0%	<span style="color: red;">■</span> Greater than 1.00%

October 2019 Version				November 2020 Version			
District	2010 Redistricting Population	Number	Percent	District	2010 Redistricting Population	Number	Percent
1	36,851	99	0.27%	1	36,851	49	0.13%
2	36,847	-18	-0.05%	2	36,847	-2	-0.01%
3	36,852	51	0.14%	3	36,852	22	0.06%
4	36,844	30	0.08%	4	36,844	55	0.15%
5	36,876	55	0.15%	5	36,876	-79	-0.21%
6	36,851	-264	-0.72%	6	36,851	-15	-0.04%
7	36,855	-24	-0.07%	7	36,855	11	0.03%
8	36,867	-19	-0.05%	8	36,867	17	0.05%
9	36,845	239	0.65%	9	36,845	27	0.07%
10	36,870	-98	-0.27%	10	36,870	-33	-0.09%
11	36,871	-462	-1.25%	11	36,871	113	0.31%
12	36,876	236	0.64%	12	36,876	29	0.08%
13	36,859	-193	-0.52%	13	36,859	-7	-0.02%
14	36,873	-166	-0.45%	14	36,873	40	0.11%
15	36,852	67	0.18%	15	36,852	5	0.01%
16	36,850	-13	-0.04%	16	36,850	-155	-0.42%
17	36,871	22	0.06%	17	36,871	78	0.21%
18	36,852	71	0.19%	18	36,852	-83	-0.23%
19	36,874	14	0.04%	19	36,874	24	0.07%
20	36,855	-35	-0.09%	20	36,855	-20	-0.05%
21	36,832	-372	-1.01%	21	36,832	-28	-0.08%
22	36,862	-112	-0.30%	22	36,862	17	0.05%
23	36,855	59	0.16%	23	36,855	7	0.02%
24	36,852	227	0.62%	24	36,852	-48	-0.13%
25	36,856	-119	-0.32%	25	36,856	-40	-0.11%
26	36,850	436	1.18%	26	36,850	92	0.25%
27	36,857	125	0.34%	27	36,857	-3	-0.01%
28	36,864	207	0.56%	28	36,864	17	0.05%
29	36,853	-394	-1.07%	29	36,853	-100	-0.27%
30	36,858	-235	-0.64%	30	36,858	-14	-0.04%
31	36,852	-71	-0.19%	31	36,852	-29	-0.08%
32	36,839	-74	-0.20%	32	36,839	71	0.19%

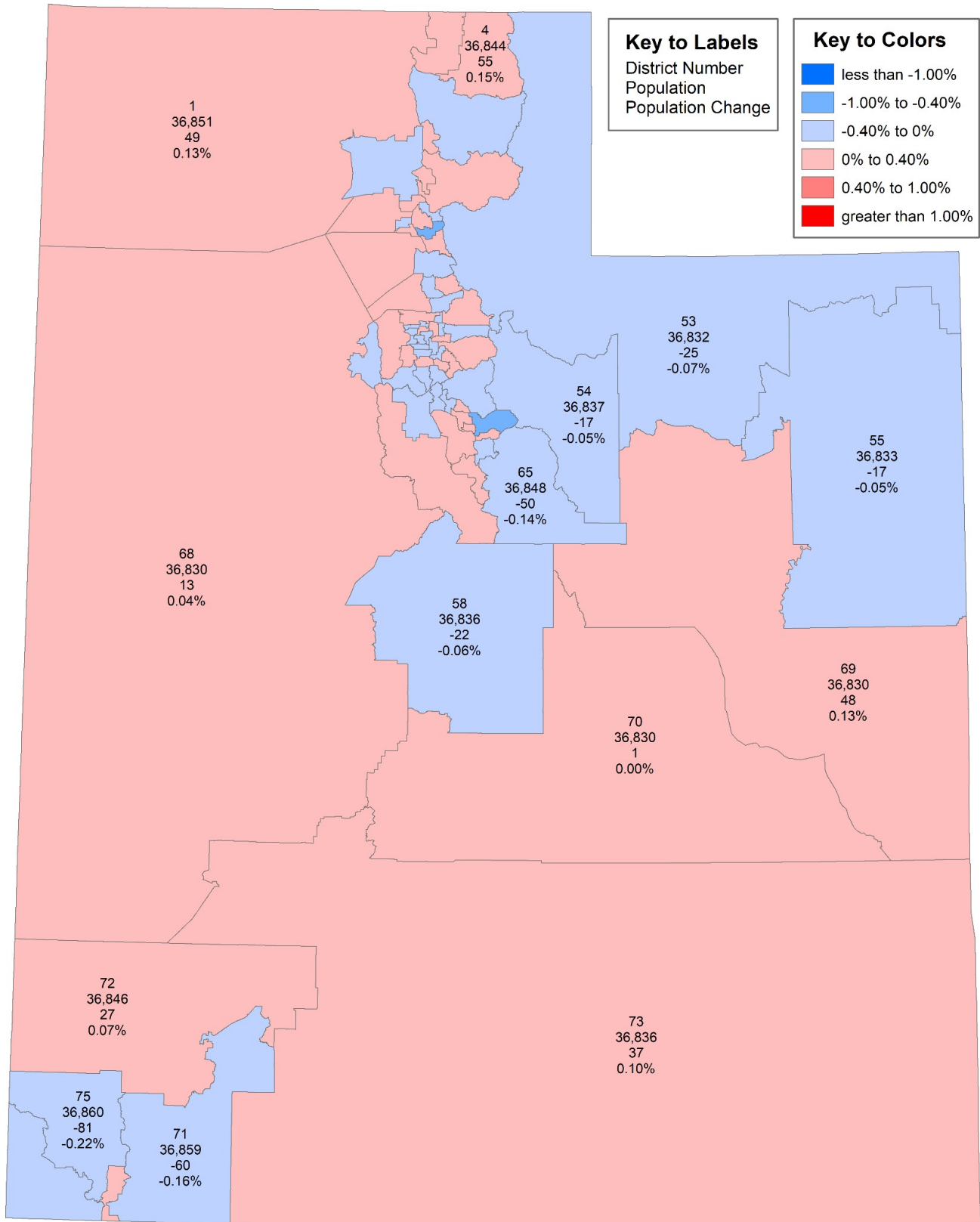


## House Districts

33	36,845	22	0.06%	33	36,845	70	0.19%
34	36,851	-159	-0.43%	34	36,851	-20	-0.05%
35	36,860	-155	-0.42%	35	36,860	23	0.06%
36	36,843	3	0.01%	36	36,843	-29	-0.08%
37	36,841	360	0.98%	37	36,841	15	0.04%
38	36,847	70	0.19%	38	36,847	-19	-0.05%
39	36,859	182	0.49%	39	36,859	-17	-0.05%
40	36,855	-265	-0.72%	40	36,855	-40	-0.11%
41	36,844	-67	-0.18%	41	36,844	-25	-0.07%
42	36,857	-75	-0.20%	42	36,857	54	0.15%
43	36,857	-386	-1.05%	43	36,857	-70	-0.19%
44	36,847	-184	-0.50%	44	36,847	-16	-0.04%
45	36,856	-146	-0.40%	45	36,856	-12	-0.03%
46	36,854	-143	-0.39%	46	36,854	35	0.09%
47	36,851	34	0.09%	47	36,851	-23	-0.06%
48	36,842	-80	-0.22%	48	36,842	-150	-0.41%
49	36,856	247	0.67%	49	36,856	30	0.08%
50	36,844	265	0.72%	50	36,844	49	0.13%
51	36,853	-143	-0.39%	51	36,853	-42	-0.11%
52	36,841	-135	-0.37%	52	36,841	-43	-0.12%
53	36,832	260	0.71%	53	36,832	-25	-0.07%
54	36,837	-44	-0.12%	54	36,837	-17	-0.05%
55	36,833	177	0.48%	55	36,833	-17	-0.05%
56	36,852	-254	-0.69%	56	36,852	-7	-0.02%
57	36,854	137	0.37%	57	36,854	41	0.11%
58	36,836	152	0.41%	58	36,836	-22	-0.06%
59	36,844	12	0.03%	59	36,844	84	0.23%
60	36,851	-117	-0.32%	60	36,851	60	0.16%
61	36,853	-3	-0.01%	61	36,853	2	0.01%
62	36,839	-91	-0.25%	62	36,839	122	0.33%
63	36,855	26	0.07%	63	36,855	47	0.13%
64	36,846	-215	-0.58%	64	36,846	-50	-0.14%
65	36,848	83	0.23%	65	36,848	-50	-0.14%
66	36,857	25	0.07%	66	36,857	52	0.14%
67	36,859	-175	-0.47%	67	36,859	14	0.04%
68	36,830	835	2.27%	68	36,830	13	0.04%
69	36,830	126	0.34%	69	36,830	48	0.13%
70	36,830	128	0.35%	70	36,830	1	0.00%
71	36,859	215	0.58%	71	36,859	-60	-0.16%
72	36,846	1	0.00%	72	36,846	27	0.07%
73	36,836	463	1.26%	73	36,836	37	0.10%
74	36,874	-382	-1.04%	74	36,874	-7	-0.02%
75	36,860	127	0.34%	75	36,860	-81	-0.22%



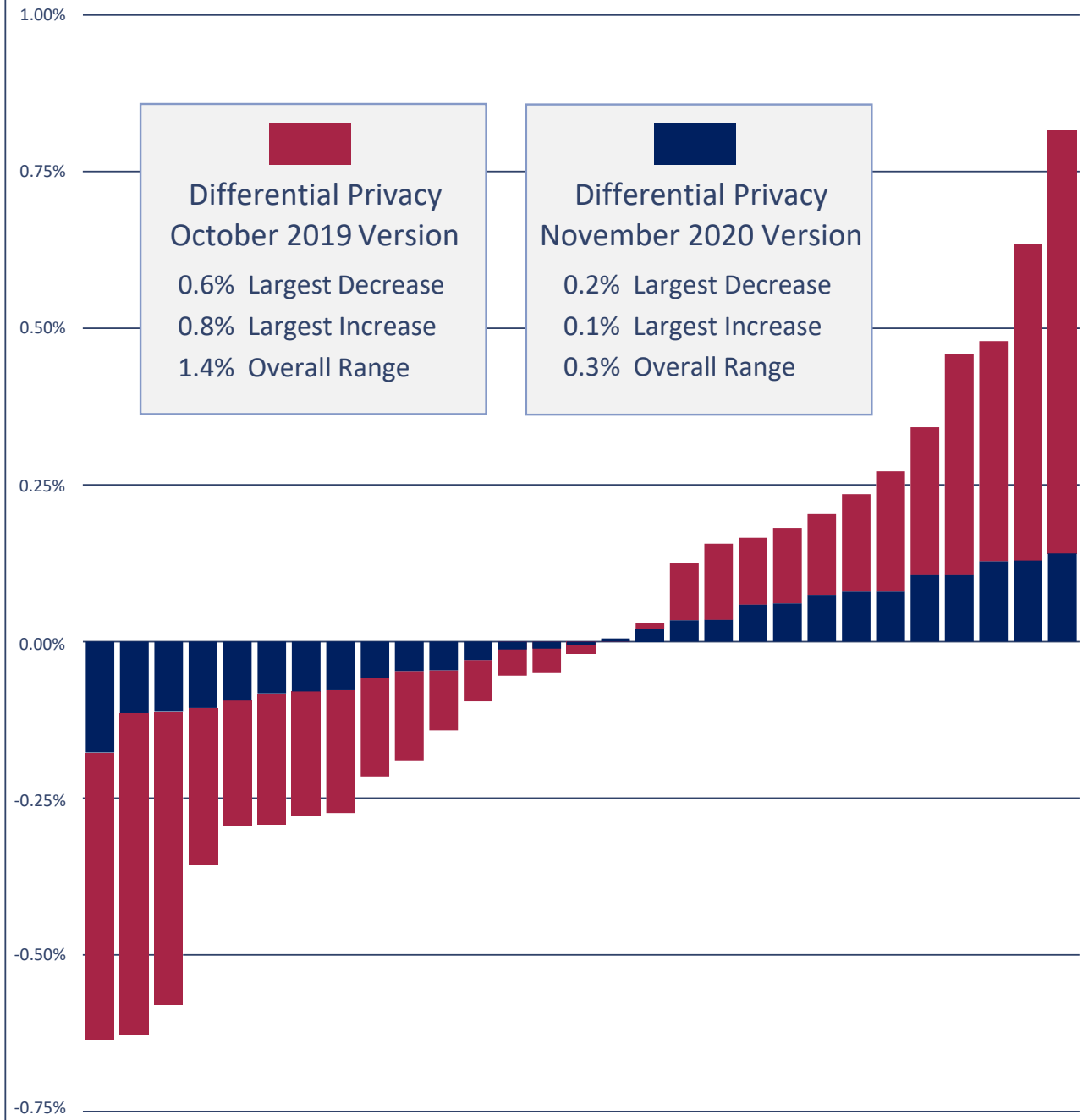
## House Districts





## Senate Districts

### Differential Privacy Applied to 2010 Senate District Populations



## Senate Districts

### Differential Privacy Applied to 2010 Senate District Populations

#### Key to Colors

<span style="color: blue;">■</span> Less than -0.40%	<span style="color: lightcoral;">■</span> 0% to 0.20%
<span style="color: lightblue;">■</span> -0.40% to -0.20%	<span style="color: coral;">■</span> 0.20% to 0.50%
<span style="color: lightblue;">■</span> -0.20% to 0%	<span style="color: red;">■</span> greater than 0.50%

October 2019 Version				November 2020 Version			
District	2010 Redistricting Population	Number	Percent	District	2010 Redistricting Population	Number	Percent
12	95,304	-605	-0.63%	19	95,309	-169	-0.18%
3	95,304	-597	-0.63%	11	95,306	-109	-0.11%
22	95,305	-552	-0.58%	18	95,307	-107	-0.11%
11	95,306	-339	-0.36%	4	95,308	-101	-0.11%
16	95,306	-280	-0.29%	25	95,305	-90	-0.09%
5	95,307	-278	-0.29%	27	95,307	-79	-0.08%
15	95,306	-266	-0.28%	16	95,306	-76	-0.08%
29	95,309	-261	-0.27%	14	95,309	-74	-0.08%
18	95,307	-205	-0.22%	9	95,306	-56	-0.06%
21	95,306	-182	-0.19%	23	95,307	-45	-0.05%
13	95,305	-135	-0.14%	12	95,304	-44	-0.05%
6	95,306	-91	-0.10%	6	95,306	-28	-0.03%
19	95,309	-52	-0.05%	7	95,306	-12	-0.01%
14	95,309	-47	-0.05%	28	95,303	-11	-0.01%
7	95,306	-18	-0.02%	5	95,307	-6	-0.01%
8	95,309	1	0.00%	29	95,309	5	0.01%
4	95,308	28	0.03%	3	95,304	19	0.02%
1	95,304	119	0.12%	2	95,308	32	0.03%
9	95,306	149	0.16%	20	95,304	33	0.03%
23	95,307	158	0.17%	1	95,304	56	0.06%
20	95,304	173	0.18%	13	95,305	58	0.06%
17	95,307	194	0.20%	21	95,306	71	0.07%
28	95,303	224	0.24%	17	95,307	76	0.08%
25	95,305	259	0.27%	24	95,307	76	0.08%
26	95,307	326	0.34%	8	95,309	101	0.11%
27	95,307	437	0.46%	22	95,305	101	0.11%
10	95,308	457	0.48%	26	95,307	122	0.13%
2	95,308	605	0.63%	10	95,308	123	0.13%
24	95,307	778	0.82%	15	95,306	134	0.14%



Senate Districts

Differential Privacy Applied to 2010 Senate District Populations

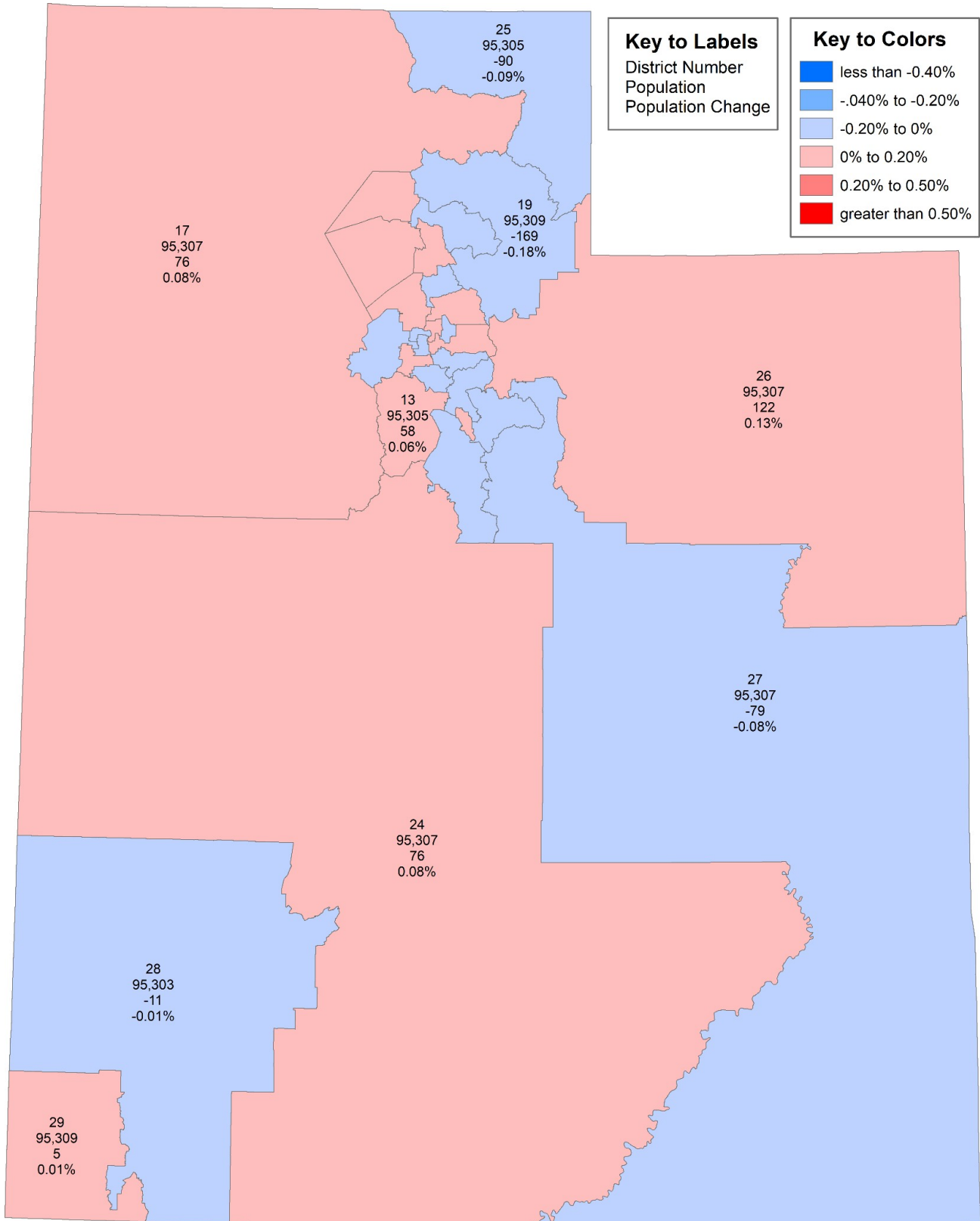
Key to Colors

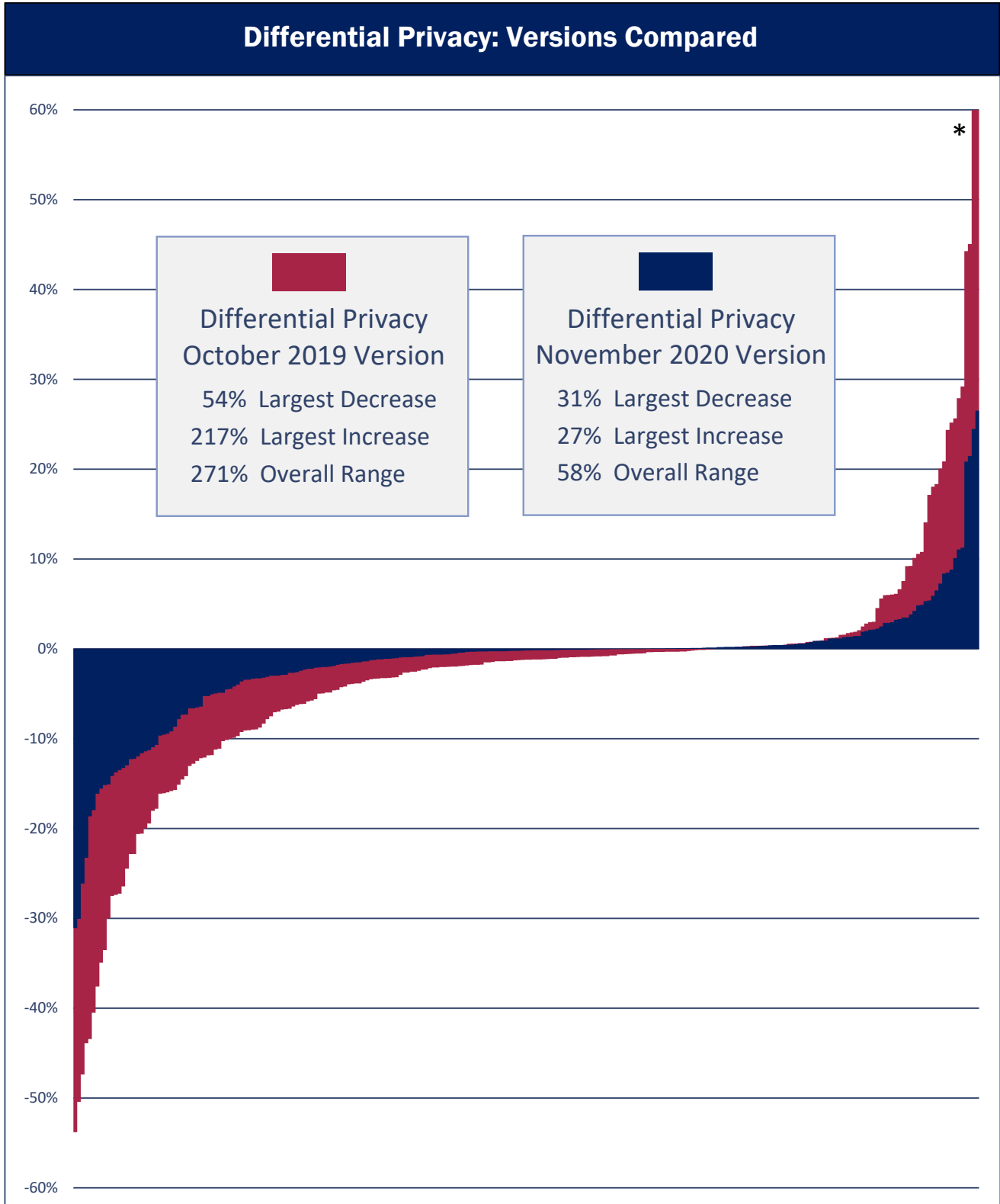
- Less than -0.40%
- .040% to -0.20%
- 0.20% to 0%
- 0% to 0.20%
- 0.20% to 0.50%
- greater than 0.50%

October 2019 Version				November 2020 Version			
District	2010 Redistricting Population	Number	Percent	District	2010 Redistricting Population	Number	Percent
1	95,304	119	0.12%	1	95,304	56	0.06%
2	95,308	605	0.63%	2	95,308	32	0.03%
3	95,304	-597	-0.63%	3	95,304	19	0.02%
4	95,308	28	0.03%	4	95,308	-101	-0.11%
5	95,307	-278	-0.29%	5	95,307	-6	-0.01%
6	95,306	-91	-0.10%	6	95,306	-28	-0.03%
7	95,306	-18	-0.02%	7	95,306	-12	-0.01%
8	95,309	1	0.00%	8	95,309	101	0.11%
9	95,306	149	0.16%	9	95,306	-56	-0.06%
10	95,308	457	0.48%	10	95,308	123	0.13%
11	95,306	-339	-0.36%	11	95,306	-109	-0.11%
12	95,304	-605	-0.63%	12	95,304	-44	-0.05%
13	95,305	-135	-0.14%	13	95,305	58	0.06%
14	95,309	-47	-0.05%	14	95,309	-74	-0.08%
15	95,306	-266	-0.28%	15	95,306	134	0.14%
16	95,306	-280	-0.29%	16	95,306	-76	-0.08%
17	95,307	194	0.20%	17	95,307	76	0.08%
18	95,307	-205	-0.22%	18	95,307	-107	-0.11%
19	95,309	-52	-0.05%	19	95,309	-169	-0.18%
20	95,304	173	0.18%	20	95,304	33	0.03%
21	95,306	-182	-0.19%	21	95,306	71	0.07%
22	95,305	-552	-0.58%	22	95,305	101	0.11%
23	95,307	158	0.17%	23	95,307	-45	-0.05%
24	95,307	778	0.82%	24	95,307	76	0.08%
25	95,305	259	0.27%	25	95,305	-90	-0.09%
26	95,307	326	0.34%	26	95,307	122	0.13%
27	95,307	437	0.46%	27	95,307	-79	-0.08%
28	95,303	224	0.24%	28	95,303	-11	-0.01%
29	95,309	-261	-0.27%	29	95,309	5	0.01%



Senate Districts





\* Outliers not shown. Largest increases were 217% and 121%. See page 18.





Municipalities

Differential Privacy Applied to 2010 Municipal Populations

Key to Colors

- less than -10%
- -10% to -3%
- -3% to 0%

- 0% to 3%
- 3% to 10%
- greater than 10%

October 2019 Version				November 2020 Version			
Municipality	2010 Census Population	Number	Percent	Municipality	2010 Census Population	Number	Percent
Tabiona	171	-92	-53.80%	Alton	119	-37	-31.09%
Alton	119	-60	-50.42%	Clawson	163	-49	-30.06%
Kingston	173	-82	-47.40%	Manila	310	-81	-26.13%
Henrieville	230	-101	-43.91%	Rockville	245	-57	-23.27%
Bryce Canyon City	198	-86	-43.43%	Bicknell	327	-61	-18.65%
Clawson	163	-66	-40.49%	Cannonville	167	-30	-17.96%
Hatch	133	-50	-37.59%	Mayfield	496	-80	-16.13%
Brian Head	83	-29	-34.94%	Woodruff	180	-28	-15.56%
Snowville	167	-56	-33.53%	Meadow	310	-47	-15.16%
Manila	310	-93	-30.00%	Randolph	464	-70	-15.09%
Sigurd	429	-118	-27.51%	Levan	841	-119	-14.15%
Hanksville	219	-60	-27.40%	Koosharem	327	-45	-13.76%
Fayette	242	-66	-27.27%	Garden City	562	-76	-13.52%
Meadow	310	-82	-26.45%	Leamington	226	-30	-13.27%
Bicknell	327	-80	-24.46%	Rush Valley	447	-58	-12.98%
Rockville	245	-56	-22.86%	Tabiona	171	-21	-12.28%
Oak City	578	-132	-22.84%	Portage	245	-30	-12.24%
Goshen	921	-190	-20.63%	Snowville	167	-20	-11.98%
Rush Valley	447	-92	-20.58%	Redmond	730	-85	-11.64%
Altamont	225	-45	-20.00%	Sigurd	429	-49	-11.42%
Spring City	988	-192	-19.43%	Hatch	133	-15	-11.28%
Amalga	488	-88	-18.03%	Castle Valley	319	-35	-10.97%
Redmond	730	-130	-17.81%	Scipio	327	-35	-10.70%
Laketown	248	-40	-16.13%	New Harmony	207	-20	-9.66%
Loa	572	-92	-16.08%	Elmo	418	-40	-9.57%
Francis	1,077	-172	-15.97%	Goshen	921	-87	-9.45%
Bear River City	853	-135	-15.83%	Hinckley	696	-64	-9.20%
Hideout	656	-103	-15.70%	Glendale	381	-33	-8.66%
Joseph	344	-52	-15.12%	Henrieville	230	-18	-7.83%
Kanosh	474	-69	-14.56%	Howell	245	-18	-7.35%
Springdale	529	-75	-14.18%	Independence	164	-12	-7.32%



# Differential Privacy

## Municipalities

New Harmony	207	-27	-13.04%	Fountain Green	1,071	-71	-6.63%
Enterprise	1,711	-219	-12.80%	Deweyville	332	-22	-6.63%
Levan	841	-105	-12.49%	Alta	383	-25	-6.53%
Escalante	797	-97	-12.17%	Springdale	529	-34	-6.43%
Fairview	1,247	-151	-12.11%	Bear River City	853	-45	-5.28%
Cleveland	464	-55	-11.85%	Ophir	38	-2	-5.26%
Sterling	262	-31	-11.83%	Charleston	415	-21	-5.06%
Randolph	464	-52	-11.21%	Mona	1,547	-77	-4.98%
Vernon	243	-27	-11.11%	Marysville	408	-20	-4.90%
Orangeville	1,470	-151	-10.27%	Altamont	225	-11	-4.89%
Plymouth	414	-42	-10.14%	Cleveland	464	-21	-4.53%
Tropic	530	-53	-10.00%	Henefer	766	-34	-4.44%
Virgin	596	-59	-9.90%	Fairfield	119	-5	-4.20%
Castle Valley	319	-31	-9.72%	Boulder	226	-9	-3.98%
Holden	378	-35	-9.26%	Oak City	578	-21	-3.63%
Midway	3,845	-350	-9.10%	Fairview	1,247	-43	-3.45%
East Carbon	1,301	-118	-9.07%	Holden	378	-13	-3.44%
Woodland Hills	1,344	-121	-9.00%	Wellington	1,676	-56	-3.34%
Ferron	1,626	-146	-8.98%	Newton	789	-26	-3.30%
Circleville	547	-48	-8.78%	Circleville	547	-18	-3.29%
Cornish	288	-24	-8.33%	Tropic	530	-17	-3.21%
Alta	383	-30	-7.83%	Huntsville	608	-19	-3.13%
Mona	1,547	-116	-7.50%	Orangeville	1,470	-44	-2.99%
Clarkston	666	-47	-7.06%	Toquerville	1,370	-41	-2.99%
Fountain Green	1,071	-75	-7.00%	Kamas	1,811	-54	-2.98%
Elk Ridge	2,436	-165	-6.77%	Millville	1,829	-53	-2.90%
Newton	789	-53	-6.72%	Midway	3,845	-111	-2.89%
Myton	569	-38	-6.68%	Paragonah	488	-13	-2.66%
Orderville	577	-37	-6.41%	Lewiston	1,766	-47	-2.66%
Mayfield	496	-31	-6.25%	River Heights	1,734	-45	-2.60%
Fielding	455	-28	-6.15%	Aurora	1,016	-25	-2.46%
Howell	245	-15	-6.12%	Leeds	820	-19	-2.32%
Manti	3,276	-192	-5.86%	East Carbon	1,301	-29	-2.23%
Moroni	1,423	-82	-5.76%	Monroe	2,256	-50	-2.22%
Hinckley	696	-39	-5.60%	Eureka	669	-14	-2.09%
Ballard	801	-40	-4.99%	Castle Dale	1,630	-34	-2.09%
Toquerville	1,370	-68	-4.96%	Panguitch	1,520	-31	-2.04%
Independence	164	-8	-4.88%	Bryce Canyon City	198	-4	-2.02%
Mount Pleasant	3,260	-159	-4.88%	Moroni	1,423	-28	-1.97%
Green River	952	-44	-4.62%	Coalville	1,363	-26	-1.91%
Kanab	4,312	-197	-4.57%	Stockton	616	-11	-1.79%
Nibley	5,438	-232	-4.27%	Gunnison	3,285	-56	-1.70%
Monticello	1,972	-83	-4.21%	Naples	1,755	-29	-1.65%
Lewiston	1,766	-70	-3.96%	Monticello	1,972	-32	-1.62%



# Differential Privacy

## Municipalities

Gunnison	3,285	-128	-3.90%	Apple Valley	701	-11	-1.57%
Farr West	5,928	-229	-3.86%	Uintah	1,322	-20	-1.51%
Fillmore	2,435	-94	-3.86%	Virgin	596	-9	-1.51%
Willard	1,772	-65	-3.67%	Milford	1,409	-20	-1.42%
Garland	2,400	-84	-3.50%	Mount Pleasant	3,260	-45	-1.38%
Coalville	1,363	-46	-3.37%	Annabella	795	-10	-1.26%
Honeyville	1,441	-48	-3.33%	Duchesne	1,690	-21	-1.24%
Naples	1,755	-58	-3.30%	Perry	4,512	-53	-1.17%
Smithfield	9,495	-310	-3.26%	Corinne	685	-8	-1.17%
Kamas	1,811	-59	-3.26%	Fillmore	2,435	-27	-1.11%
Glenwood	464	-15	-3.23%	Elk Ridge	2,436	-27	-1.11%
Elsinore	847	-27	-3.19%	Daniel	938	-10	-1.07%
Moab	5,046	-160	-3.17%	Morgan	3,687	-38	-1.03%
Plain City	5,476	-159	-2.90%	Manti	3,276	-32	-0.98%
Minersville	907	-24	-2.65%	Kanab	4,312	-41	-0.95%
Payson	18,294	-484	-2.65%	Richmond	2,470	-23	-0.93%
Morgan	3,687	-94	-2.55%	Garland	2,400	-22	-0.92%
Eureka	669	-17	-2.54%	Wendover	1,400	-12	-0.86%
Aurora	1,016	-25	-2.46%	Woods Cross	9,761	-82	-0.84%
Ephraim	6,135	-143	-2.33%	Antimony	122	-1	-0.82%
Cedar City	28,857	-667	-2.31%	Paradise	904	-6	-0.66%
Beaver	3,112	-67	-2.15%	Richfield	7,551	-50	-0.66%
Henefer	766	-16	-2.09%	Blanding	3,375	-22	-0.65%
Wellington	1,676	-35	-2.09%	Enterprise	1,711	-11	-0.64%
Portage	245	-5	-2.04%	Pleasant View	7,979	-51	-0.64%
Castle Dale	1,630	-33	-2.02%	Kanosh	474	-3	-0.63%
Price	8,715	-175	-2.01%	Sunset	5,122	-32	-0.62%
Centerfield	1,367	-27	-1.98%	Torrey	182	-1	-0.55%
Ivins	6,753	-133	-1.97%	Woodland Hills	1,344	-7	-0.52%
Huntington	2,129	-41	-1.93%	Mapleton	7,979	-34	-0.43%
Oakley	1,470	-28	-1.90%	Vernon	243	-1	-0.41%
Paragonah	488	-9	-1.84%	Midvale	27,964	-101	-0.36%
Deweyville	332	-6	-1.81%	Grantsville	8,893	-29	-0.33%
Garden City	562	-10	-1.78%	Clearfield	30,112	-95	-0.32%
Duchesne	1,690	-30	-1.78%	Tremonton	7,647	-23	-0.30%
Salina	2,489	-44	-1.77%	Ephraim	6,135	-18	-0.29%
Pleasant View	7,979	-120	-1.50%	Enoch	5,803	-17	-0.29%
Santa Clara	6,003	-90	-1.50%	Providence	7,075	-20	-0.28%
Alpine	9,555	-140	-1.47%	Centerville	15,335	-43	-0.28%
Spanish Fork	34,691	-482	-1.39%	Francis	1,077	-3	-0.28%
Roosevelt	6,046	-84	-1.39%	Heber	11,362	-30	-0.26%
Cedar Hills	9,796	-135	-1.38%	Riverton	38,753	-102	-0.26%
Monroe	2,256	-31	-1.37%	Highland	15,523	-39	-0.25%
South Ogden	16,532	-225	-1.36%	Willard	1,772	-4	-0.23%



Municipalities

North Logan	8,269	-107	-1.29%	Smithfield	9,495	-21	-0.22%
Tremonton	7,647	-97	-1.27%	Parowan	2,790	-6	-0.22%
Clearfield	30,112	-375	-1.25%	Cedar City	28,857	-56	-0.19%
Saratoga Springs	17,781	-218	-1.23%	North Ogden	17,357	-31	-0.18%
Marysvale	408	-5	-1.23%	Nibley	5,438	-9	-0.17%
Enoch	5,803	-69	-1.19%	Harrisville	5,567	-9	-0.16%
Heber	11,362	-134	-1.18%	Moab	5,046	-8	-0.16%
Woods Cross	9,761	-115	-1.18%	Eagle Mountain	21,415	-33	-0.15%
Mendon	1,282	-15	-1.17%	Farmington	18,275	-28	-0.15%
Brigham City	17,899	-206	-1.15%	Lehi	47,407	-72	-0.15%
Kanarrville	355	-4	-1.13%	Cedar Hills	9,796	-14	-0.14%
Park City	7,558	-85	-1.12%	Kaysville	27,300	-35	-0.13%
Perry	4,512	-46	-1.02%	Washington Terrace	9,067	-11	-0.12%
Roy	36,884	-370	-1.00%	Minersville	907	-1	-0.11%
Apple Valley	701	-7	-1.00%	Syracuse	24,331	-25	-0.10%
South Salt Lake	23,617	-224	-0.95%	Springville	29,466	-29	-0.10%
Lehi	47,407	-448	-0.95%	Hooper	7,218	-7	-0.10%
Wendover	1,400	-13	-0.93%	South Salt Lake	23,617	-22	-0.09%
Providence	7,075	-64	-0.90%	Clinton	20,426	-19	-0.09%
Hyrum	7,609	-68	-0.89%	Hyrum	7,609	-7	-0.09%
Sandy	87,461	-776	-0.89%	St. George	72,897	-65	-0.09%
Syracuse	24,331	-213	-0.88%	Tooele	31,605	-24	-0.08%
Midvale	27,964	-239	-0.85%	Draper	42,274	-27	-0.06%
La Verkin	4,060	-34	-0.84%	South Jordan	50,418	-32	-0.06%
Farmington	18,275	-151	-0.83%	Huntington	2,129	-1	-0.05%
American Fork	26,263	-213	-0.81%	North Salt Lake	16,322	-7	-0.04%
Kaysville	27,300	-205	-0.75%	Lincoln	10,070	-3	-0.03%
Richfield	7,551	-56	-0.74%	West Jordan	103,712	-25	-0.02%
Leeds	820	-5	-0.61%	Orem	88,328	-21	-0.02%
Riverton	38,753	-236	-0.61%	Taylorsville	58,652	-12	-0.02%
Hildale	2,726	-16	-0.59%	Salem	6,423	-1	-0.02%
Richmond	2,470	-14	-0.57%	West Valley City	129,480	-20	-0.02%
Tooele	31,605	-174	-0.55%	Ivins	6,753	-1	-0.01%
North Salt Lake	16,322	-86	-0.53%	Ogden	82,825	-11	-0.01%
Riverdale	8,426	-43	-0.51%	American Fork	26,263	-1	0.00%
Clinton	20,426	-100	-0.49%	Layton	67,311	-1	0.00%
Draper	42,274	-161	-0.38%	Lyman	258	0	0.00%
Provo	112,488	-417	-0.37%	Provo	112,488	2	0.00%
West Haven	10,272	-37	-0.36%	Spanish Fork	34,691	3	0.01%
Harrisville	5,567	-20	-0.36%	Price	8,715	1	0.01%
Washington Terrace	9,067	-31	-0.34%	Salt Lake City	186,440	29	0.02%
Hurricane	13,748	-46	-0.33%	Farr West	5,928	2	0.03%
Cottonwood Heights	33,433	-107	-0.32%	Bountiful	42,552	15	0.04%
Orem	88,328	-279	-0.32%	Pleasant Grove	33,509	15	0.04%



# Municipalities

West Valley City	129,480	-408	-0.32%	Roy	36,884	25	0.07%
Pleasant Grove	33,509	-104	-0.31%	Murray	46,746	33	0.07%
West Jordan	103,712	-313	-0.30%	Payson	18,294	13	0.07%
Logan	48,174	-117	-0.24%	Holladay	26,472	20	0.08%
Eagle Mountain	21,415	-44	-0.21%	Logan	48,174	43	0.09%
Parowan	2,790	-4	-0.14%	Fruit Heights	4,987	5	0.10%
Lindon	10,070	-14	-0.14%	Alpine	9,555	10	0.10%
St. George	72,897	-88	-0.12%	Cottonwood Heights	33,433	36	0.11%
Blanding	3,375	-3	-0.09%	Saratoga Springs	17,781	24	0.13%
Ogden	82,825	-62	-0.07%	Washington	18,761	27	0.14%
South Jordan	50,418	-16	-0.03%	Santa Clara	6,003	9	0.15%
Paradise	904	0	0.00%	Brigham City	17,899	34	0.19%
Bluffdale	7,598	7	0.09%	Vernal	9,089	18	0.20%
Springville	29,466	28	0.10%	Genola	1,370	3	0.22%
Taylorsville	58,652	59	0.10%	Hildale	2,726	6	0.22%
Layton	67,311	70	0.10%	Riverdale	8,426	19	0.23%
Holladay	26,472	40	0.15%	Roosevelt	6,046	14	0.23%
Hyde Park	3,833	8	0.21%	West Haven	10,272	24	0.23%
North Ogden	17,357	49	0.28%	Plain City	5,476	13	0.24%
Bountiful	42,552	126	0.30%	Delta	3,436	9	0.26%
Herriman	21,785	74	0.34%	Bluffdale	7,598	20	0.26%
Washington	18,761	64	0.34%	South Ogden	16,532	44	0.27%
Wellsville	3,432	12	0.35%	Hurricane	13,748	45	0.33%
Santaquin	9,128	32	0.35%	Sandy	87,461	297	0.34%
West Point	9,511	35	0.37%	Herriman	21,785	76	0.35%
Salt Lake City	186,440	697	0.37%	Mendon	1,282	5	0.39%
Annabella	795	3	0.38%	Nephi	5,389	22	0.41%
Big Water	475	2	0.42%	Hyde Park	3,833	16	0.42%
Centerville	15,335	65	0.42%	Green River	952	4	0.42%
South Weber	6,051	27	0.45%	North Logan	8,269	37	0.45%
Murray	46,746	262	0.56%	Hideout	656	3	0.46%
Genola	1,370	8	0.58%	West Point	9,511	44	0.46%
Vernal	9,089	54	0.59%	Park City	7,558	41	0.54%
Milford	1,409	9	0.64%	Salina	2,489	14	0.56%
Sunset	5,122	33	0.64%	La Verkin	4,060	24	0.59%
Nephi	5,389	41	0.76%	Santaquin	9,128	61	0.67%
Panguitch	1,520	12	0.79%	Helper	2,201	16	0.73%
Huntsville	608	5	0.82%	Mantua	687	6	0.87%
Leamington	226	2	0.88%	Loa	572	5	0.87%
Highland	15,523	144	0.93%	Honeyville	1,441	13	0.90%
West Bountiful	5,265	62	1.18%	West Bountiful	5,265	48	0.91%
Grantsville	8,893	108	1.21%	South Weber	6,051	69	1.14%
Salem	6,423	79	1.23%	Kingston	173	2	1.16%
Helper	2,201	28	1.27%	Elwood	1,034	12	1.16%



River Heights	1,734	27	1.56%	Joseph	344	4	1.16%
Delta	3,436	54	1.57%	Oakley	1,470	19	1.29%
Hooper	7,218	127	1.76%	Beaver	3,112	42	1.35%
Mapleton	7,979	146	1.83%	Clarkston	666	9	1.35%
Mantua	687	13	1.89%	Spring City	988	14	1.42%
Millville	1,829	38	2.08%	Amalga	488	7	1.43%
Corinne	685	17	2.48%	Wellsville	3,432	65	1.89%
Elwood	1,034	29	2.80%	Centerfield	1,367	27	1.98%
Fruit Heights	4,987	148	2.97%	Elsinore	847	18	2.13%
Marriott-Slaterville	1,701	51	3.00%	Trenton	464	10	2.16%
Uintah	1,322	60	4.54%	Marriott-Slaterville	1,701	38	2.23%
Rocky Ridge	733	41	5.59%	Ballard	801	20	2.50%
Elmo	418	25	5.98%	Escalante	797	23	2.89%
Cannonville	167	10	5.99%	Fayette	242	7	2.89%
Glendale	381	23	6.04%	Ferron	1,626	48	2.95%
Woodruff	180	11	6.11%	Central Valley	528	17	3.22%
Sunnyside	377	25	6.63%	Fielding	455	15	3.30%
Lynndyl	106	8	7.55%	Cornish	288	10	3.47%
Wallsburg	250	23	9.20%	Emery	288	10	3.47%
Cedar Fort	368	34	9.24%	Rocky Ridge	733	28	3.82%
Charleston	415	42	10.12%	Kanarrville	355	15	4.23%
Daniel	938	99	10.55%	Plymouth	414	20	4.83%
Central Valley	528	57	10.80%	Cedar Fort	368	18	4.89%
Koosharem	327	46	14.07%	Sunnyside	377	20	5.31%
Scipio	327	56	17.13%	Glenwood	464	25	5.39%
Emery	288	52	18.06%	Orderville	577	34	5.89%
Stockton	616	113	18.34%	Myton	569	37	6.50%
Trenton	464	93	20.04%	Laketown	248	18	7.26%
Wales	302	63	20.86%	Junction	191	16	8.38%
Fairfield	119	29	24.37%	Lynndyl	106	9	8.49%
Vineyard	139	35	25.18%	Wallsburg	250	22	8.80%
Junction	191	49	25.65%	Big Water	475	48	10.11%
Lyman	258	72	27.91%	Sterling	262	29	11.07%
Boulder	226	66	29.20%	Wales	302	34	11.26%
Antimony	122	54	44.26%	Scofield	24	5	20.83%
Torrey	182	82	45.05%	Hanksville	219	47	21.46%
Ophir	38	46	121.05%	Vineyard	139	34	24.46%
Scofield	24	52	216.67%	Brian Head	83	22	26.51%



## Municipalities

### Differential Privacy Applied to 2010 City/Town Populations

#### Key to Colors

- less than -10%
- -10% to -3%
- -3% to 0%

- 0% to 3%
- 3% to 10%
- greater than 10%

#### October 2019 Version

#### November 2020 Version

City/Town	2010 Census Population	Number	Percent	City/Town	2010 Census Population	Number	Percent
Alpine	9,555	-140	-1.47%	Alpine	9,555	10	0.10%
Alta	383	-30	-7.83%	Alta	383	-25	-6.53%
Altamont	225	-45	-20.00%	Altamont	225	-11	-4.89%
Alton	119	-60	-50.42%	Alton	119	-37	-31.09%
Amalga	488	-88	-18.03%	Amalga	488	7	1.43%
American Fork	26,263	-213	-0.81%	American Fork	26,263	-1	0.00%
Annabella	795	3	0.38%	Annabella	795	-10	-1.26%
Antimony	122	54	44.26%	Antimony	122	-1	-0.82%
Apple Valley	701	-7	-1.00%	Apple Valley	701	-11	-1.57%
Aurora	1,016	-25	-2.46%	Aurora	1,016	-25	-2.46%
Ballard	801	-40	-4.99%	Ballard	801	20	2.50%
Bear River City	853	-135	-15.83%	Bear River City	853	-45	-5.28%
Beaver	3,112	-67	-2.15%	Beaver	3,112	42	1.35%
Bicknell	327	-80	-24.46%	Bicknell	327	-61	-18.65%
Big Water	475	2	0.42%	Big Water	475	48	10.11%
Blanding	3,375	-3	-0.09%	Blanding	3,375	-22	-0.65%
Bluffdale	7,598	7	0.09%	Bluffdale	7,598	20	0.26%
Boulder	226	66	29.20%	Boulder	226	-9	-3.98%
Bountiful	42,552	126	0.30%	Bountiful	42,552	15	0.04%
Brian Head	83	-29	-34.94%	Brian Head	83	22	26.51%
Brigham City	17,899	-206	-1.15%	Brigham City	17,899	34	0.19%
Bryce Canyon City	198	-86	-43.43%	Bryce Canyon City	198	-4	-2.02%
Cannonville	167	10	5.99%	Cannonville	167	-30	-17.96%
Castle Dale	1,630	-33	-2.02%	Castle Dale	1,630	-34	-2.09%
Castle Valley	319	-31	-9.72%	Castle Valley	319	-35	-10.97%
Cedar City	28,857	-667	-2.31%	Cedar City	28,857	-56	-0.19%
Cedar Fort	368	34	9.24%	Cedar Fort	368	18	4.89%
Cedar Hills	9,796	-135	-1.38%	Cedar Hills	9,796	-14	-0.14%
Centerfield	1,367	-27	-1.98%	Centerfield	1,367	27	1.98%



Centerville	15,335	65	0.42%	Centerville	15,335	-43	-0.28%
Central Valley	528	57	10.80%	Central Valley	528	17	3.22%
Charleston	415	42	10.12%	Charleston	415	-21	-5.06%
Circleville	547	-48	-8.78%	Circleville	547	-18	-3.29%
Clarkston	666	-47	-7.06%	Clarkston	666	9	1.35%
Clawson	163	-66	-40.49%	Clawson	163	-49	-30.06%
Clearfield	30,112	-375	-1.25%	Clearfield	30,112	-95	-0.32%
Cleveland	464	-55	-11.85%	Cleveland	464	-21	-4.53%
Clinton	20,426	-100	-0.49%	Clinton	20,426	-19	-0.09%
Coalville	1,363	-46	-3.37%	Coalville	1,363	-26	-1.91%
Corinne	685	17	2.48%	Corinne	685	-8	-1.17%
Cornish	288	-24	-8.33%	Cornish	288	10	3.47%
Cottonwood Heights	33,433	-107	-0.32%	Cottonwood Heights	33,433	36	0.11%
Daniel	938	99	10.55%	Daniel	938	-10	-1.07%
Delta	3,436	54	1.57%	Delta	3,436	9	0.26%
Deweyville	332	-6	-1.81%	Deweyville	332	-22	-6.63%
Draper	42,274	-161	-0.38%	Draper	42,274	-27	-0.06%
Duchesne	1,690	-30	-1.78%	Duchesne	1,690	-21	-1.24%
Eagle Mountain	21,415	-44	-0.21%	Eagle Mountain	21,415	-33	-0.15%
East Carbon	1,301	-118	-9.07%	East Carbon	1,301	-29	-2.23%
Elk Ridge	2,436	-165	-6.77%	Elk Ridge	2,436	-27	-1.11%
Elmo	418	25	5.98%	Elmo	418	-40	-9.57%
Elsinore	847	-27	-3.19%	Elsinore	847	18	2.13%
Elwood	1,034	29	2.80%	Elwood	1,034	12	1.16%
Emery	288	52	18.06%	Emery	288	10	3.47%
Enoch	5,803	-69	-1.19%	Enoch	5,803	-17	-0.29%
Enterprise	1,711	-219	-12.80%	Enterprise	1,711	-11	-0.64%
Ephraim	6,135	-143	-2.33%	Ephraim	6,135	-18	-0.29%
Escalante	797	-97	-12.17%	Escalante	797	23	2.89%
Eureka	669	-17	-2.54%	Eureka	669	-14	-2.09%
Fairfield	119	29	24.37%	Fairfield	119	-5	-4.20%
Fairview	1,247	-151	-12.11%	Fairview	1,247	-43	-3.45%
Farmington	18,275	-151	-0.83%	Farmington	18,275	-28	-0.15%
Farr West	5,928	-229	-3.86%	Farr West	5,928	2	0.03%
Fayette	242	-66	-27.27%	Fayette	242	7	2.89%
Ferron	1,626	-146	-8.98%	Ferron	1,626	48	2.95%
Fielding	455	-28	-6.15%	Fielding	455	15	3.30%
Fillmore	2,435	-94	-3.86%	Fillmore	2,435	-27	-1.11%
Fountain Green	1,071	-75	-7.00%	Fountain Green	1,071	-71	-6.63%
Francis	1,077	-172	-15.97%	Francis	1,077	-3	-0.28%
Fruit Heights	4,987	148	2.97%	Fruit Heights	4,987	5	0.10%
Garden City	562	-10	-1.78%	Garden City	562	-76	-13.52%
Garland	2,400	-84	-3.50%	Garland	2,400	-22	-0.92%
Genola	1,370	8	0.58%	Genola	1,370	3	0.22%





# Differential Privacy

# 21

## Municipalities

Glendale	381	23	6.04%	Glendale	381	-33	-8.66%
Glenwood	464	-15	-3.23%	Glenwood	464	25	5.39%
Goshen	921	-190	-20.63%	Goshen	921	-87	-9.45%
Grantsville	8,893	108	1.21%	Grantsville	8,893	-29	-0.33%
Green River	952	-44	-4.62%	Green River	952	4	0.42%
Gunnison	3,285	-128	-3.90%	Gunnison	3,285	-56	-1.70%
Hanksville	219	-60	-27.40%	Hanksville	219	47	21.46%
Harrisville	5,567	-20	-0.36%	Harrisville	5,567	-9	-0.16%
Hatch	133	-50	-37.59%	Hatch	133	-15	-11.28%
Heber	11,362	-134	-1.18%	Heber	11,362	-30	-0.26%
Helper	2,201	28	1.27%	Helper	2,201	16	0.73%
Henefer	766	-16	-2.09%	Henefer	766	-34	-4.44%
Henrieville	230	-101	-43.91%	Henrieville	230	-18	-7.83%
Herriman	21,785	74	0.34%	Herriman	21,785	76	0.35%
Hideout	656	-103	-15.70%	Hideout	656	3	0.46%
Highland	15,523	144	0.93%	Highland	15,523	-39	-0.25%
Hildale	2,726	-16	-0.59%	Hildale	2,726	6	0.22%
Hinckley	696	-39	-5.60%	Hinckley	696	-64	-9.20%
Holden	378	-35	-9.26%	Holden	378	-13	-3.44%
Holladay	26,472	40	0.15%	Holladay	26,472	20	0.08%
Honeyville	1,441	-48	-3.33%	Honeyville	1,441	13	0.90%
Hooper	7,218	127	1.76%	Hooper	7,218	-7	-0.10%
Howell	245	-15	-6.12%	Howell	245	-18	-7.35%
Huntington	2,129	-41	-1.93%	Huntington	2,129	-1	-0.05%
Huntsville	608	5	0.82%	Huntsville	608	-19	-3.13%
Hurricane	13,748	-46	-0.33%	Hurricane	13,748	45	0.33%
Hyde Park	3,833	8	0.21%	Hyde Park	3,833	16	0.42%
Hyrum	7,609	-68	-0.89%	Hyrum	7,609	-7	-0.09%
Independence	164	-8	-4.88%	Independence	164	-12	-7.32%
Ivins	6,753	-133	-1.97%	Ivins	6,753	-1	-0.01%
Joseph	344	-52	-15.12%	Joseph	344	4	1.16%
Junction	191	49	25.65%	Junction	191	16	8.38%
Kamas	1,811	-59	-3.26%	Kamas	1,811	-54	-2.98%
Kanab	4,312	-197	-4.57%	Kanab	4,312	-41	-0.95%
Kanarraville	355	-4	-1.13%	Kanarraville	355	15	4.23%
Kanosh	474	-69	-14.56%	Kanosh	474	-3	-0.63%
Kaysville	27,300	-205	-0.75%	Kaysville	27,300	-35	-0.13%
Kingston	173	-82	-47.40%	Kingston	173	2	1.16%
Koosharem	327	46	14.07%	Koosharem	327	-45	-13.76%
La Verkin	4,060	-34	-0.84%	La Verkin	4,060	24	0.59%
Laketown	248	-40	-16.13%	Laketown	248	18	7.26%
Layton	67,311	70	0.10%	Layton	67,311	-1	0.00%
Leamington	226	2	0.88%	Leamington	226	-30	-13.27%
Leeds	820	-5	-0.61%	Leeds	820	-19	-2.32%

## Differential Privacy

22

## Municipalities

Lehi	47,407	-448	-0.95%	Lehi	47,407	-72	-0.15%
Levan	841	-105	-12.49%	Levan	841	-119	-14.15%
Lewiston	1,766	-70	-3.96%	Lewiston	1,766	-47	-2.66%
Lindon	10,070	-14	-0.14%	Lindon	10,070	-3	-0.03%
Loa	572	-92	-16.08%	Loa	572	5	0.87%
Logan	48,174	-117	-0.24%	Logan	48,174	43	0.09%
Lyman	258	72	27.91%	Lyman	258	0	0.00%
Lynndyl	106	8	7.55%	Lynndyl	106	9	8.49%
Manila	310	-93	-30.00%	Manila	310	-81	-26.13%
Manti	3,276	-192	-5.86%	Manti	3,276	-32	-0.98%
Mantua	687	13	1.89%	Mantua	687	6	0.87%
Mapleton	7,979	146	1.83%	Mapleton	7,979	-34	-0.43%
Marriott-Slaterville	1,701	51	3.00%	Marriott-Slaterville	1,701	38	2.23%
Marysville	408	-5	-1.23%	Marysville	408	-20	-4.90%
Mayfield	496	-31	-6.25%	Mayfield	496	-80	-16.13%
Meadow	310	-82	-26.45%	Meadow	310	-47	-15.16%
Mendon	1,282	-15	-1.17%	Mendon	1,282	5	0.39%
Midvale	27,964	-239	-0.85%	Midvale	27,964	-101	-0.36%
Midway	3,845	-350	-9.10%	Midway	3,845	-111	-2.89%
Milford	1,409	9	0.64%	Milford	1,409	-20	-1.42%
Millville	1,829	38	2.08%	Millville	1,829	-53	-2.90%
Minersville	907	-24	-2.65%	Minersville	907	-1	-0.11%
Moab	5,046	-160	-3.17%	Moab	5,046	-8	-0.16%
Mona	1,547	-116	-7.50%	Mona	1,547	-77	-4.98%
Monroe	2,256	-31	-1.37%	Monroe	2,256	-50	-2.22%
Monticello	1,972	-83	-4.21%	Monticello	1,972	-32	-1.62%
Morgan	3,687	-94	-2.55%	Morgan	3,687	-38	-1.03%
Moroni	1,423	-82	-5.76%	Moroni	1,423	-28	-1.97%
Mount Pleasant	3,260	-159	-4.88%	Mount Pleasant	3,260	-45	-1.38%
Murray	46,746	262	0.56%	Murray	46,746	33	0.07%
Myton	569	-38	-6.68%	Myton	569	37	6.50%
Naples	1,755	-58	-3.30%	Naples	1,755	-29	-1.65%
Nephi	5,389	41	0.76%	Nephi	5,389	22	0.41%
New Harmony	207	-27	-13.04%	New Harmony	207	-20	-9.66%
Newton	789	-53	-6.72%	Newton	789	-26	-3.30%
Nibley	5,438	-232	-4.27%	Nibley	5,438	-9	-0.17%
North Logan	8,269	-107	-1.29%	North Logan	8,269	37	0.45%
North Ogden	17,357	49	0.28%	North Ogden	17,357	-31	-0.18%
North Salt Lake	16,322	-86	-0.53%	North Salt Lake	16,322	-7	-0.04%
Oak City	578	-132	-22.84%	Oak City	578	-21	-3.63%
Oakley	1,470	-28	-1.90%	Oakley	1,470	19	1.29%
Ogden	82,825	-62	-0.07%	Ogden	82,825	-11	-0.01%
Ophir	38	46	121.05%	Ophir	38	-2	-5.26%
Orangeville	1,470	-151	-10.27%	Orangeville	1,470	-44	-2.99%



# Differential Privacy

# 23

## Municipalities

Orderville	577	-37	-6.41%	Orderville	577	34	5.89%
Orem	88,328	-279	-0.32%	Orem	88,328	-21	-0.02%
Panguitch	1,520	12	0.79%	Panguitch	1,520	-31	-2.04%
Paradise	904	0	0.00%	Paradise	904	-6	-0.66%
Paragonah	488	-9	-1.84%	Paragonah	488	-13	-2.66%
Park City	7,558	-85	-1.12%	Park City	7,558	41	0.54%
Parowan	2,790	-4	-0.14%	Parowan	2,790	-6	-0.22%
Payson	18,294	-484	-2.65%	Payson	18,294	13	0.07%
Perry	4,512	-46	-1.02%	Perry	4,512	-53	-1.17%
Plain City	5,476	-159	-2.90%	Plain City	5,476	13	0.24%
Pleasant Grove	33,509	-104	-0.31%	Pleasant Grove	33,509	15	0.04%
Pleasant View	7,979	-120	-1.50%	Pleasant View	7,979	-51	-0.64%
Plymouth	414	-42	-10.14%	Plymouth	414	20	4.83%
Portage	245	-5	-2.04%	Portage	245	-30	-12.24%
Price	8,715	-175	-2.01%	Price	8,715	1	0.01%
Providence	7,075	-64	-0.90%	Providence	7,075	-20	-0.28%
Provo	112,488	-417	-0.37%	Provo	112,488	2	0.00%
Randolph	464	-52	-11.21%	Randolph	464	-70	-15.09%
Redmond	730	-130	-17.81%	Redmond	730	-85	-11.64%
Richfield	7,551	-56	-0.74%	Richfield	7,551	-50	-0.66%
Richmond	2,470	-14	-0.57%	Richmond	2,470	-23	-0.93%
River Heights	1,734	27	1.56%	River Heights	1,734	-45	-2.60%
Riverdale	8,426	-43	-0.51%	Riverdale	8,426	19	0.23%
Riverton	38,753	-236	-0.61%	Riverton	38,753	-102	-0.26%
Rockville	245	-56	-22.86%	Rockville	245	-57	-23.27%
Rocky Ridge	733	41	5.59%	Rocky Ridge	733	28	3.82%
Roosevelt	6,046	-84	-1.39%	Roosevelt	6,046	14	0.23%
Roy	36,884	-370	-1.00%	Roy	36,884	25	0.07%
Rush Valley	447	-92	-20.58%	Rush Valley	447	-58	-12.98%
Salem	6,423	79	1.23%	Salem	6,423	-1	-0.02%
Salina	2,489	-44	-1.77%	Salina	2,489	14	0.56%
Salt Lake City	186,440	697	0.37%	Salt Lake City	186,440	29	0.02%
Sandy	87,461	-776	-0.89%	Sandy	87,461	297	0.34%
Santa Clara	6,003	-90	-1.50%	Santa Clara	6,003	9	0.15%
Santaquin	9,128	32	0.35%	Santaquin	9,128	61	0.67%
Saratoga Springs	17,781	-218	-1.23%	Saratoga Springs	17,781	24	0.13%
Scipio	327	56	17.13%	Scipio	327	-35	-10.70%
Scofield	24	52	216.67%	Scofield	24	5	20.83%
Sigurd	429	-118	-27.51%	Sigurd	429	-49	-11.42%
Smithfield	9,495	-310	-3.26%	Smithfield	9,495	-21	-0.22%
Snowville	167	-56	-33.53%	Snowville	167	-20	-11.98%
South Jordan	50,418	-16	-0.03%	South Jordan	50,418	-32	-0.06%
South Ogden	16,532	-225	-1.36%	South Ogden	16,532	44	0.27%
South Salt Lake	23,617	-224	-0.95%	South Salt Lake	23,617	-22	-0.09%

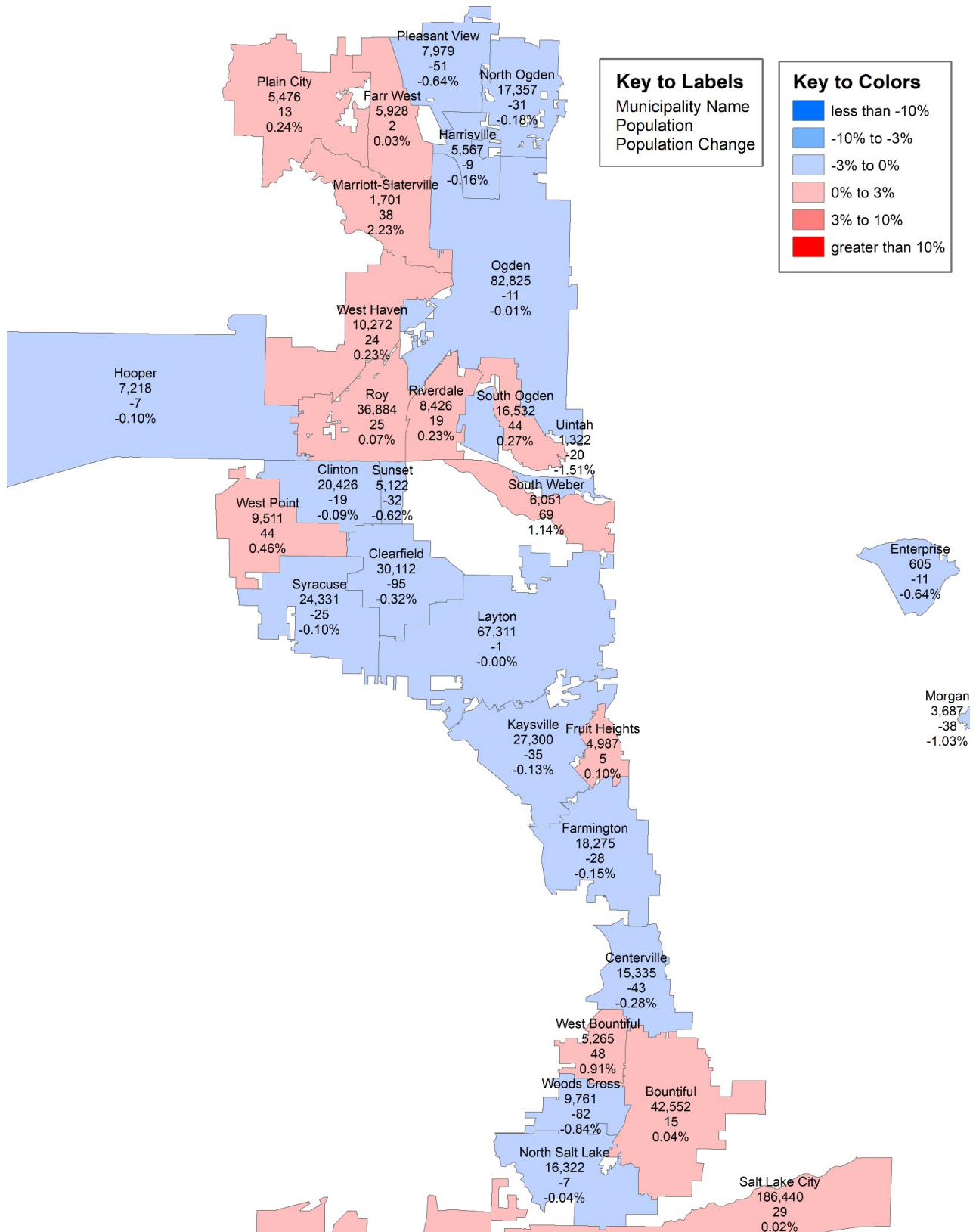


## Municipalities

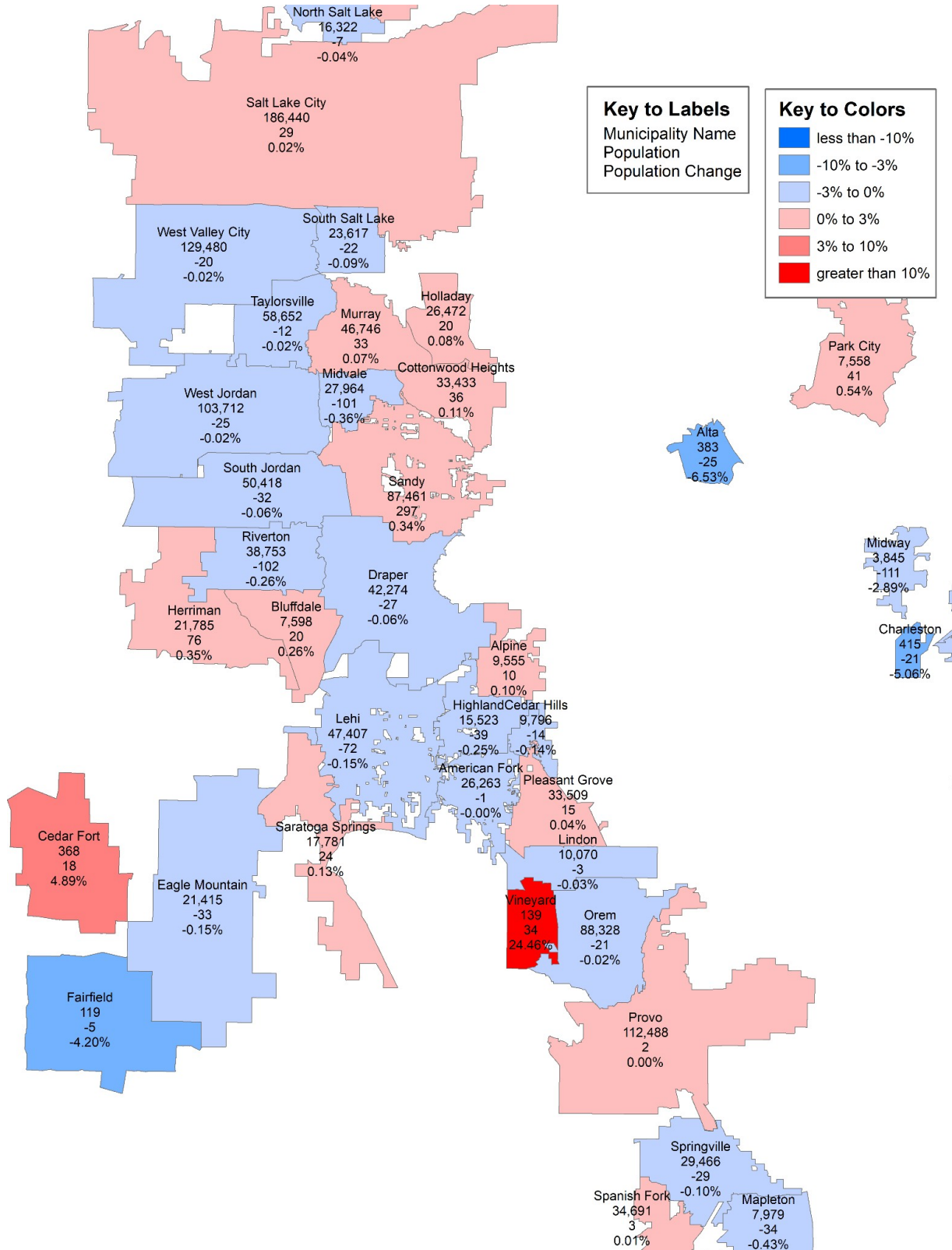
South Weber	6,051	27	0.45%	South Weber	6,051	69	1.14%
Spanish Fork	34,691	-482	-1.39%	Spanish Fork	34,691	3	0.01%
Spring City	988	-192	-19.43%	Spring City	988	14	1.42%
Springdale	529	-75	-14.18%	Springdale	529	-34	-6.43%
Springville	29,466	28	0.10%	Springville	29,466	-29	-0.10%
St. George	72,897	-88	-0.12%	St. George	72,897	-65	-0.09%
Sterling	262	-31	-11.83%	Sterling	262	29	11.07%
Stockton	616	113	18.34%	Stockton	616	-11	-1.79%
Sunnyside	377	25	6.63%	Sunnyside	377	20	5.31%
Sunset	5,122	33	0.64%	Sunset	5,122	-32	-0.62%
Syracuse	24,331	-213	-0.88%	Syracuse	24,331	-25	-0.10%
Tabiona	171	-92	-53.80%	Tabiona	171	-21	-12.28%
Taylorville	58,652	59	0.10%	Taylorville	58,652	-12	-0.02%
Tooele	31,605	-174	-0.55%	Tooele	31,605	-24	-0.08%
Toquerville	1,370	-68	-4.96%	Toquerville	1,370	-41	-2.99%
Torrey	182	82	45.05%	Torrey	182	-1	-0.55%
Tremonton	7,647	-97	-1.27%	Tremonton	7,647	-23	-0.30%
Trenton	464	93	20.04%	Trenton	464	10	2.16%
Tropic	530	-53	-10.00%	Tropic	530	-17	-3.21%
Uintah	1,322	60	4.54%	Uintah	1,322	-20	-1.51%
Vernal	9,089	54	0.59%	Vernal	9,089	18	0.20%
Vernon	243	-27	-11.11%	Vernon	243	-1	-0.41%
Vineyard	139	35	25.18%	Vineyard	139	34	24.46%
Virgin	596	-59	-9.90%	Virgin	596	-9	-1.51%
Wales	302	63	20.86%	Wales	302	34	11.26%
Wallsburg	250	23	9.20%	Wallsburg	250	22	8.80%
Washington	18,761	64	0.34%	Washington	18,761	27	0.14%
Washington Terrace	9,067	-31	-0.34%	Washington Terrace	9,067	-11	-0.12%
Wellington	1,676	-35	-2.09%	Wellington	1,676	-56	-3.34%
Wellsville	3,432	12	0.35%	Wellsville	3,432	65	1.89%
Wendover	1,400	-13	-0.93%	Wendover	1,400	-12	-0.86%
West Bountiful	5,265	62	1.18%	West Bountiful	5,265	48	0.91%
West Haven	10,272	-37	-0.36%	West Haven	10,272	24	0.23%
West Jordan	103,712	-313	-0.30%	West Jordan	103,712	-25	-0.02%
West Point	9,511	35	0.37%	West Point	9,511	44	0.46%
West Valley City	129,480	-408	-0.32%	West Valley City	129,480	-20	-0.02%
Willard	1,772	-65	-3.67%	Willard	1,772	-4	-0.23%
Woodland Hills	1,344	-121	-9.00%	Woodland Hills	1,344	-7	-0.52%
Woodruff	180	11	6.11%	Woodruff	180	-28	-15.56%
Woods Cross	9,761	-115	-1.18%	Woods Cross	9,761	-82	-0.84%



## Select Municipalities

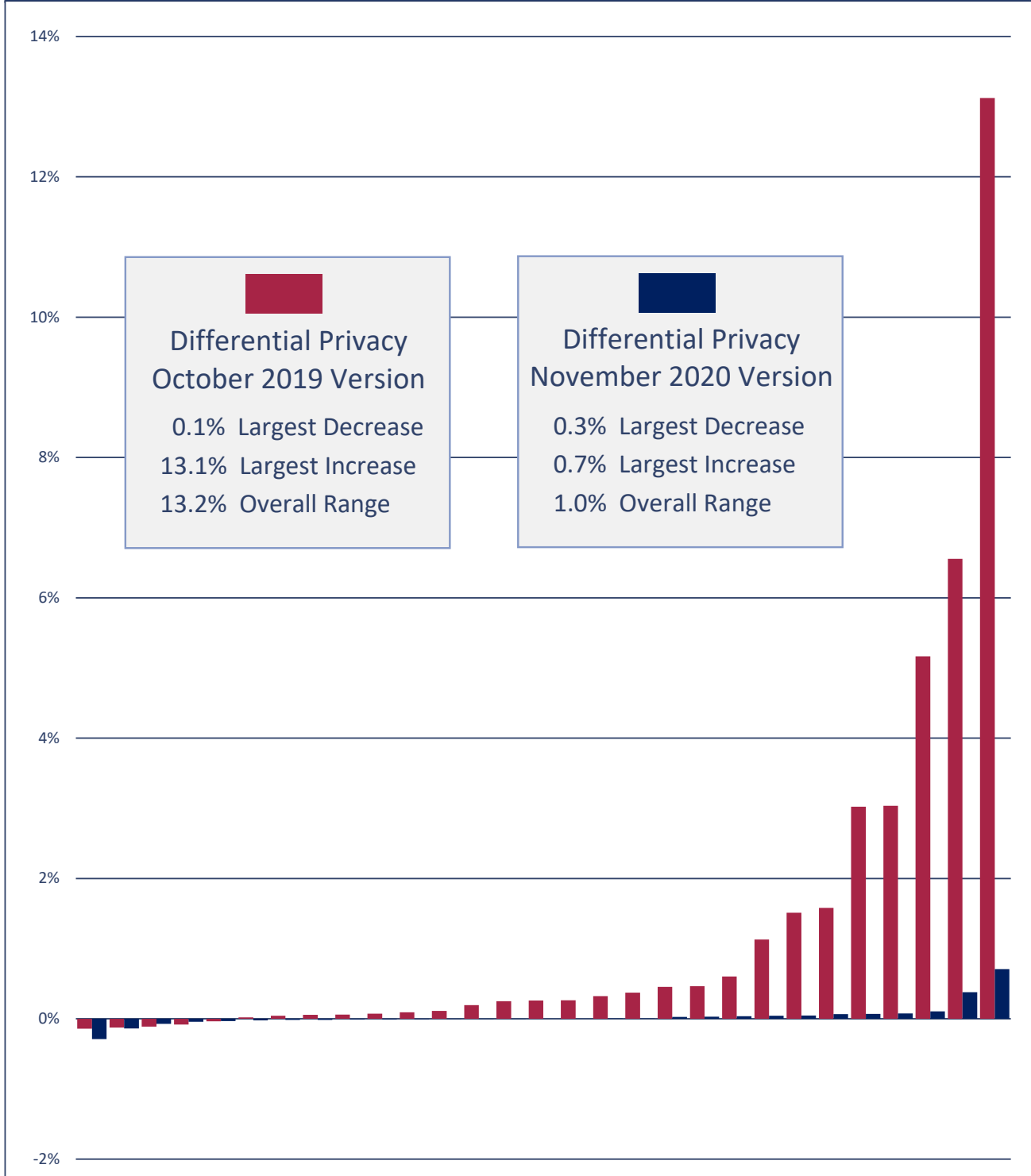


### Select Municipalities











### Differential Privacy Applied to 2010 County Populations



**Counties****Differential Privacy Applied to 2010 County Populations****Key to Colors**

 less than -0.12%  
 -0.12% to -0.05%  
 -0.05% to 0%

 0% to 1.00%  
 1.00% to 7.00%  
 greater than 7.00%

October 2019 Version				November 2020 Version			
County	2010 Redistricting Population	Number	Percent	County	2010 Redistricting Population	Number	Percent
Davis	306,479	-433	-0.14%	Wayne	2,778	-8	-0.29%
Washington	138,115	-173	-0.13%	Morgan	9,469	-13	-0.14%
Utah	516,564	-570	-0.11%	Kane	7,125	-5	-0.07%
Weber	231,236	-191	-0.08%	Sevier	20,802	-9	-0.04%
Salt Lake	1,029,655	-374	-0.04%	Carbon	21,403	-7	-0.03%
Tooele	58,218	11	0.02%	Grand	9,225	-2	-0.02%
Box Elder	49,975	22	0.04%	Uintah	32,588	-5	-0.02%
Sanpete	27,822	16	0.06%	Sanpete	27,822	-4	-0.01%
Uintah	32,588	20	0.06%	Cache	112,656	-11	-0.01%
Cache	112,656	82	0.07%	Washington	138,115	-13	-0.01%
Iron	46,163	43	0.09%	Summit	36,324	-2	-0.01%
Summit	36,324	41	0.11%	Davis	306,479	-7	0.00%
Kane	7,125	14	0.20%	Weber	231,236	-2	0.00%
San Juan	14,746	37	0.25%	Salt Lake	1,029,655	-1	0.00%
Grand	9,225	24	0.26%	Duchesne	18,607	0	0.00%
Wasatch	23,530	62	0.26%	Emery	10,976	0	0.00%
Duchesne	18,607	60	0.32%	Utah	516,564	1	0.00%
Carbon	21,403	80	0.37%	Tooele	58,218	2	0.00%
Emery	10,976	50	0.46%	Box Elder	49,975	13	0.03%
Sevier	20,802	97	0.47%	Iron	46,163	14	0.03%
Morgan	9,469	57	0.60%	Wasatch	23,530	9	0.04%
Beaver	6,629	75	1.13%	Rich	2,264	1	0.04%
Millard	12,503	189	1.51%	Millard	12,503	6	0.05%
Juab	10,246	162	1.58%	San Juan	14,746	10	0.07%
Wayne	2,778	84	3.02%	Juab	10,246	7	0.07%
Garfield	5,172	157	3.04%	Garfield	5,172	4	0.08%
Rich	2,264	117	5.17%	Beaver	6,629	7	0.11%
Piute	1,556	102	6.56%	Daggett	1,059	4	0.38%
Daggett	1,059	139	13.13%	Piute	1,556	11	0.71%





Counties

Differential Privacy Applied to 2010 County Populations

Key to Colors

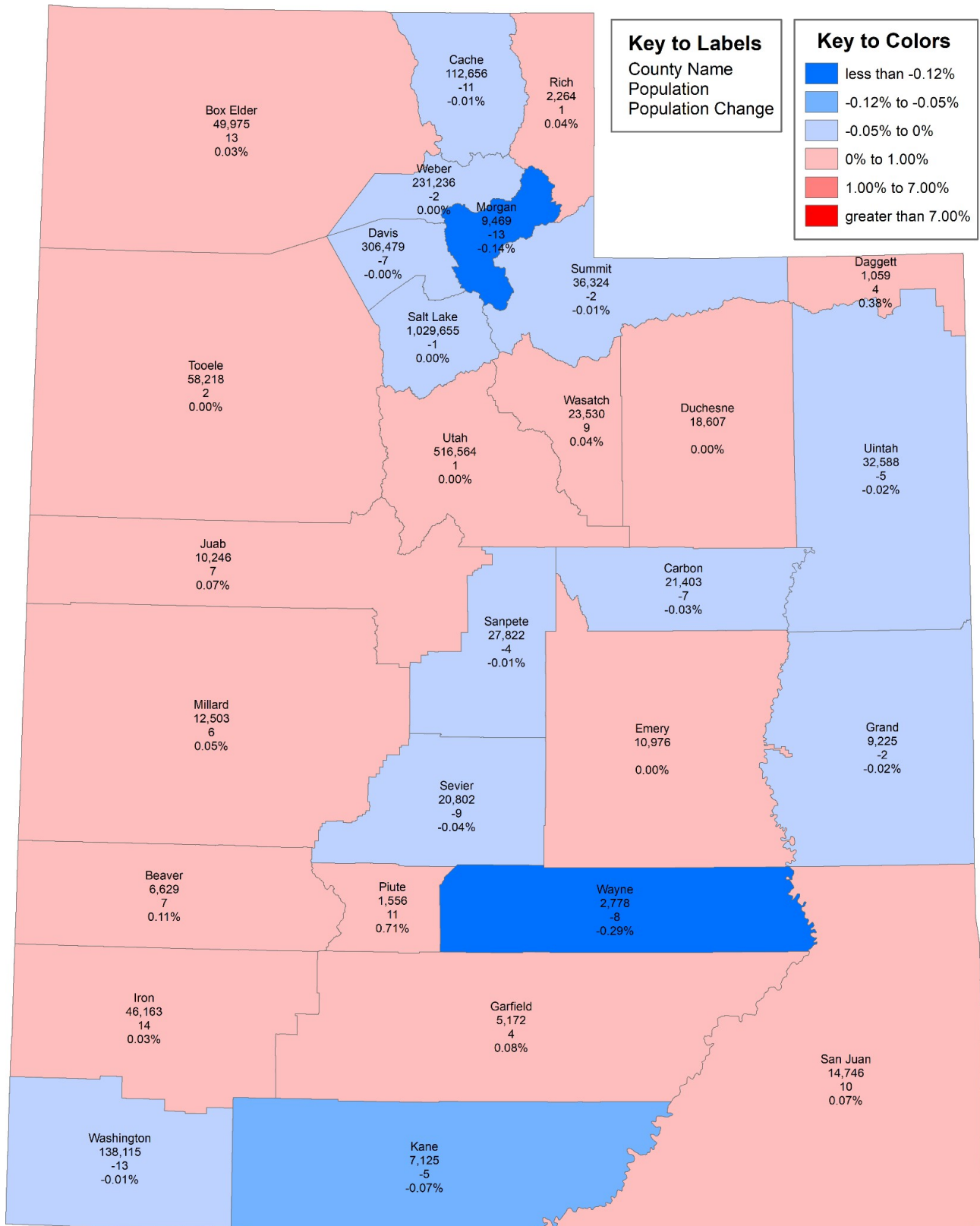
- less than -0.12%
- -0.12% to -0.05%
- -0.05% to 0%
- 0% to 1.00%
- 1.00% to 7.00%
- greater than 7.00%

October 2019 Version				November 2020 Version			
County	2010 Redistricting Population	Number	Percent	County	2010 Redistricting Population	Number	Percent
Beaver	6,629	75	1.13%	Beaver	6,629	7	0.11%
Box Elder	49,975	22	0.04%	Box Elder	49,975	13	0.03%
Cache	112,656	82	0.07%	Cache	112,656	-11	-0.01%
Carbon	21,403	80	0.37%	Carbon	21,403	-7	-0.03%
Daggett	1,059	139	13.13%	Daggett	1,059	4	0.38%
Davis	306,479	-433	-0.14%	Davis	306,479	-7	0.00%
Duchesne	18,607	60	0.32%	Duchesne	18,607	0	0.00%
Emery	10,976	50	0.46%	Emery	10,976	0	0.00%
Garfield	5,172	157	3.04%	Garfield	5,172	4	0.08%
Grand	9,225	24	0.26%	Grand	9,225	-2	-0.02%
Iron	46,163	43	0.09%	Iron	46,163	14	0.03%
Juab	10,246	162	1.58%	Juab	10,246	7	0.07%
Kane	7,125	14	0.20%	Kane	7,125	-5	-0.07%
Millard	12,503	189	1.51%	Millard	12,503	6	0.05%
Morgan	9,469	57	0.60%	Morgan	9,469	-13	-0.14%
Piute	1,556	102	6.56%	Piute	1,556	11	0.71%
Rich	2,264	117	5.17%	Rich	2,264	1	0.04%
Salt Lake	1,029,655	-374	-0.04%	Salt Lake	1,029,655	-1	0.00%
San Juan	14,746	37	0.25%	San Juan	14,746	10	0.07%
Sanpete	27,822	16	0.06%	Sanpete	27,822	-4	-0.01%
Sevier	20,802	97	0.47%	Sevier	20,802	-9	-0.04%
Summit	36,324	41	0.11%	Summit	36,324	-2	-0.01%
Tooele	58,218	11	0.02%	Tooele	58,218	2	0.00%
Uintah	32,588	20	0.06%	Uintah	32,588	-5	-0.02%
Utah	516,564	-570	-0.11%	Utah	516,564	1	0.00%
Wasatch	23,530	62	0.26%	Wasatch	23,530	9	0.04%
Washington	138,115	-173	-0.13%	Washington	138,115	-13	-0.01%
Wayne	2,778	84	3.02%	Wayne	2,778	-8	-0.29%
Weber	231,236	-191	-0.08%	Weber	231,236	-2	0.00%

# Differential Privacy

# 30

## Counties

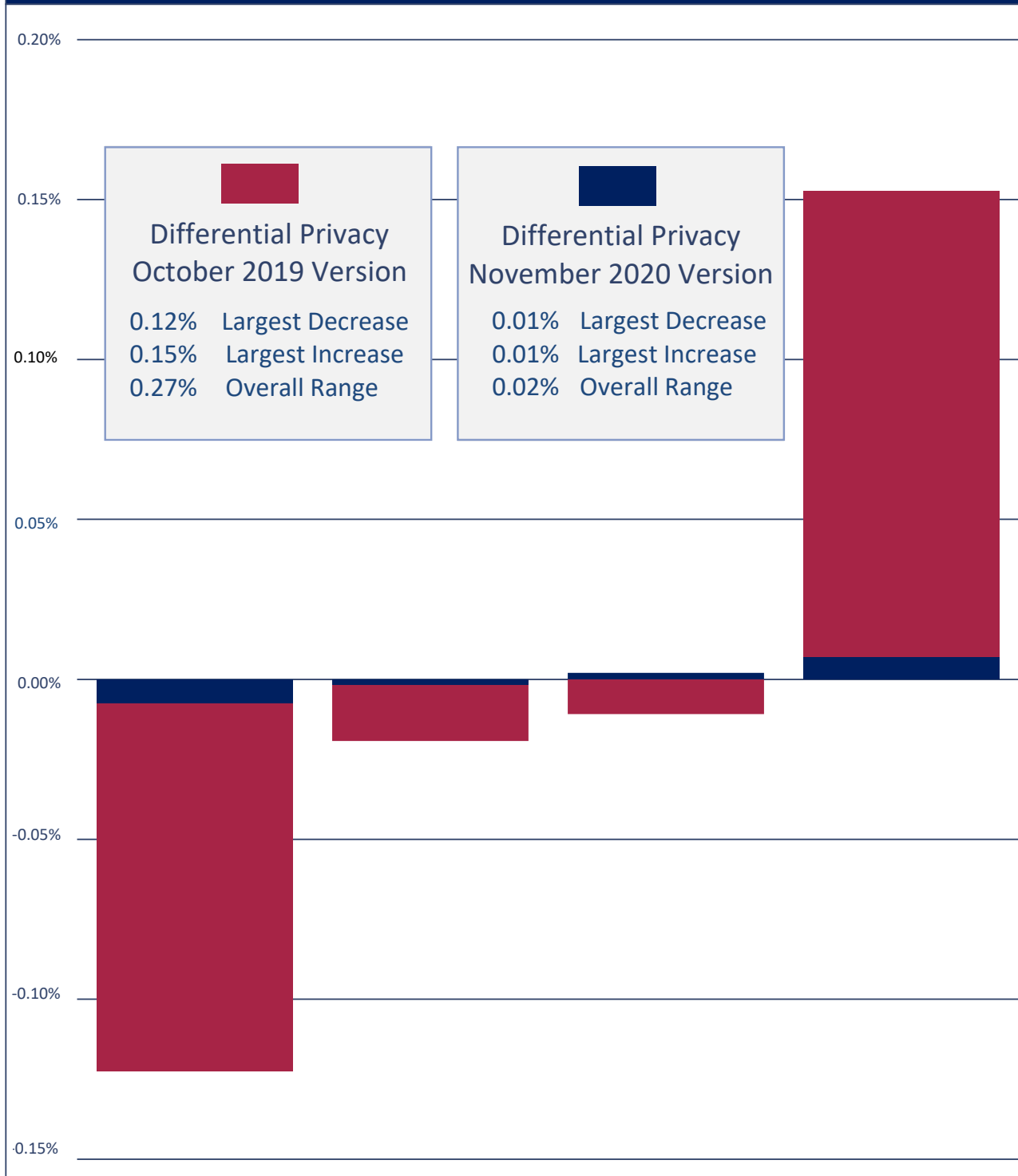




# Differential Privacy

## Congressional Districts

### Differential Privacy Applied to 2010 Congressional District Populations





## Congressional Districts

### Differential Privacy Applied to 2010 Congressional District Populations

#### Key to Colors

<span style="color: blue;">■</span> Less than -0.100%	<span style="color: lightcoral;">■</span> 0.000% to 0.020%
<span style="color: lightblue;">■</span> -0.100% to -0.020%	<span style="color: coral;">■</span> 0.020% to 0.100%
<span style="color: lightgrey;">■</span> -0.020% to 0.000%	<span style="color: red;">■</span> greater than 0.100%

October 2019 Version				November 2020 Version			
District	2010 Redistricting Population	Number	Percent	District	2010 Redistricting Population	Number	Percent
4	690,971	-846	-0.122%	4	690,971	-50	-0.007%
3	690,972	-133	-0.019%	2	690,971	-12	-0.002%
1	690,971	-75	-0.011%	1	690,971	14	0.002%
2	690,971	1054	0.153%	3	690,972	48	0.007%

### Differential Privacy Applied to 2010 Congressional District Populations

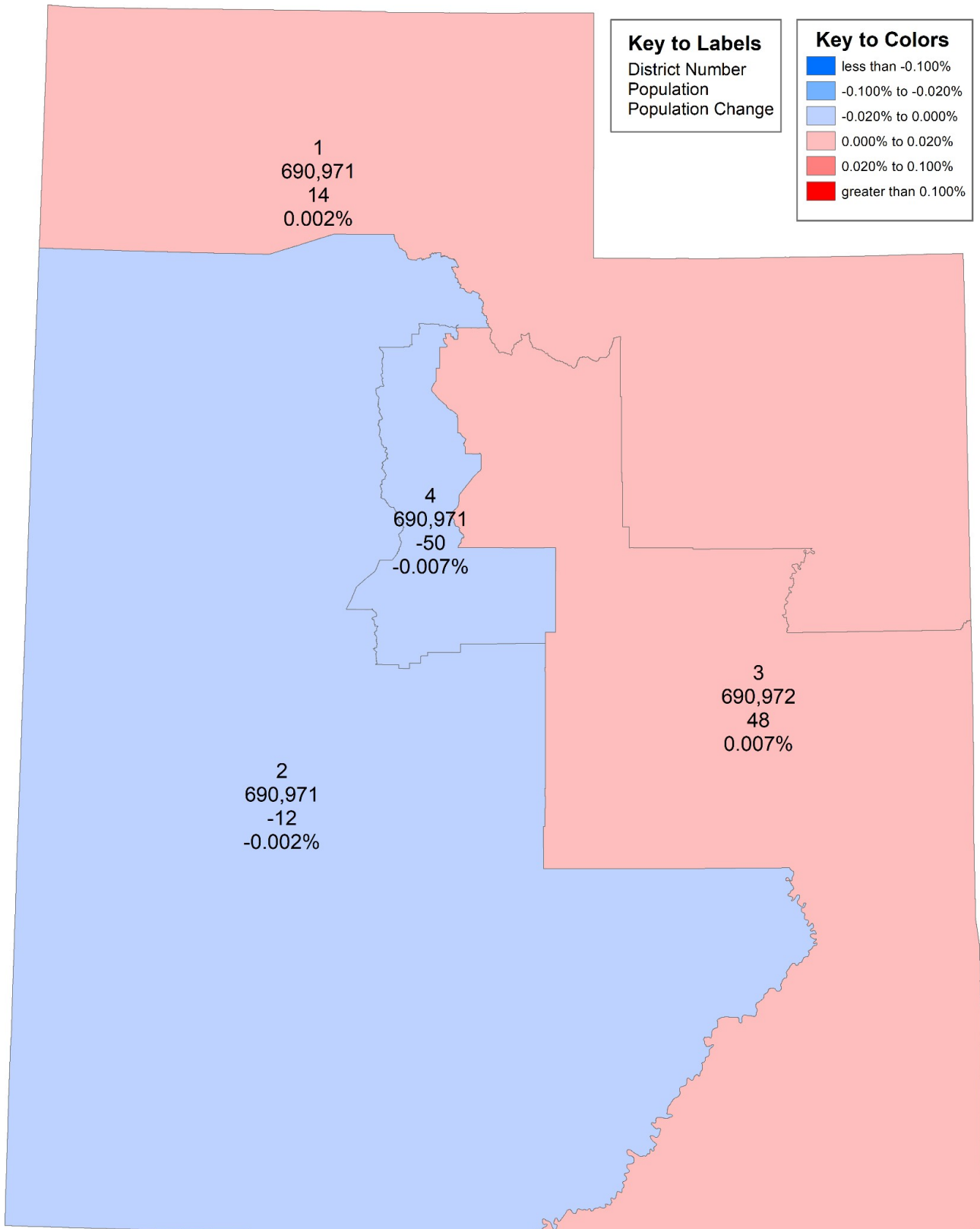
#### Key to Colors

<span style="color: blue;">■</span> Less than -0.100%	<span style="color: lightcoral;">■</span> 0.000% to 0.020%
<span style="color: lightblue;">■</span> -0.100% to -0.020%	<span style="color: coral;">■</span> 0.020% to 0.100%
<span style="color: lightgrey;">■</span> -0.020% to 0.000%	<span style="color: red;">■</span> greater than 0.100%

October 2019 Version				November 2020 Version			
District	2010 Redistricting Population	Number	Percent	District	2010 Redistricting Population	Number	Percent
1	690,971	-75	-0.011%	1	690,971	14	0.002%
2	690,971	1054	0.153%	2	690,971	-12	-0.002%
3	690,972	-133	-0.019%	3	690,972	48	0.007%
4	690,971	-846	-0.122%	4	690,971	-50	-0.007%



## Congressional Districts



# **The Effect of the Differential Privacy Disclosure Avoidance System Proposed by the Census Bureau on 2020 Census Products: Four Case Studies of Census Blocks in Alaska.**

David A. Swanson, University of California Riverside, Riverside, CA and  
The Center for Studies in Demography and Ecology, University of Washington, Seattle, WA  
(email: [dswanson@ucr.edu](mailto:dswanson@ucr.edu))

T. M. Bryan, Bryan Demographic Research, Richmond, VA  
(email: [tom@bryangeodemo.com](mailto:tom@bryangeodemo.com))

Richard Sewell, Alaska Department of Transportation and Public Facilities, Anchorage, AK  
(email: [resewell@gmail.com](mailto:resewell@gmail.com))

## **Abstract**

The Census Bureau plans to introduce a new Disclosure Avoidance System known as Differential Privacy (DP) for its 2020 census data products. Using two DP demonstration product files provided by the Census Bureau, we assess the errors introduced by DP on census block data in Alaska in the form of four case studies and find them to be substantial by type and level. We use both the May 27<sup>th</sup> 2020 DP demonstration product and the most recent, the April 28<sup>th</sup> 2021 DP demonstration product relative to our four cases studies and compare the changes. This comparison is important because the Census Bureau reports that accuracy should improve because the privacy budget was increased in response to user complaints about poor accuracy. We find that the April 28<sup>th</sup>, 2021 release does produce more accurate data but that the level of accuracy remains unsuitable for use by those who work with small area data. Because it is likely that the results we found in Alaska will be found in other states, our examination leads us to conclude that it is likely that the errors introduced by DP of the type and at the level found in the most recent demonstration product file we examined will render the nation's block level data essentially unusable.

## **Introduction**

The Census Bureau plans to introduce a new Disclosure Avoidance System known as Differential Privacy (DP) for its 2020 census data products (Abowd, 2020, Census Bureau 2020a, 2020b, 202c, 2020d, 2020e, 2020f, and 2020g). Our purpose in this paper is to assess the errors introduced by (DP) on census block data in Alaska in the form of four case studies.

Ruggles et al. (2019: 406) argue that DP goes far beyond what is necessary to keep data safe under census law and precedent and because it focuses on concealing individual characteristics instead of respondent identities, DP is a blunt and inefficient instrument for disclosure control. They go on to note that because the core metric of DP does not measure the risk of identity disclosure, it cannot assess disclosure risk as defined under census law, making it untenable for optimizing the privacy/usability trade-off.

## **Background**

Covering 570,641 square miles of land, Alaska is the largest state but with the 2010 census showing only 710,231 people, it is the least densely-populated of the 50 states, at 1.24 people per square mile (Hunsinger et al. 2012: 8). The 2010 census (see below), organized the state into 45,292 census blocks, of which only 12,870 had one or more people, leaving 32,422 without any population. On average, there were 15.68 persons in each of these 45,292 census blocks. If we look at the 12,870 census blocks with at least one person, there were 55.2 persons on average in each of these 12,870 blocks. These summary

statistics make Alaska one of the states in which one would expect a high level of disclosure avoidance at the block level because there are so few people on average per block. This is a point to which we return in the final section.

## Data

The application of DP is a brand new approach for the Census Bureau and is different from all prior Census Bureau initiatives in regard to disclosure avoidance. As a component of the DP initiative, the Census Bureau has released a series of “demonstration products” (Abowd, 2020, Census Bureau 2020a, 2020b, 2020c, 2020d, 2020e, 2020f, and 2020g) that allow outside analysts and stakeholders to determine for their purposes the impact DP would have on Census data. These demonstration products generally contain:

- the most common, basic demographic and housing variables
- different levels of geography
- data as they were originally reported in the Summary Files (SF) in 2010, which reported actual census data with small privacy protection modifications as noted supra page
- trial data as they have been by adjusted (perturbed) DP

As the Census Bureau responded to User complaints about poor accuracy, the “privacy budgets” were changed in the demonstration products to provide higher levels of accuracy (Beveridge, 2021). Here, we examine the errors introduced by DP on 2010 Census block data for Alaska in the form of four case studies. In our initial analysis, we employ the “demonstration product” for census blocks in Alaska released May 27th, 2020, file (labeled as 2020527) with an epsilon level of 4.0, which was downloaded from the Minnesota Population Center’s NHGIS site: <https://nhgis.org/privacy-protected-demonstration-data>. Against the results we find from the May 27th, 2020 file, we compare results from the most recent release, April 28th, 2021. (file labeled as 20210428) with an epsilon level of 10.3, which was downloaded from the same site.

In the analyses for case studies 1 through 3, we employed the cross-tabulation routine found in Release 12 of the NCSST Statistical System (<https://www.ncss.com/software/ncss/>). For case study 4, we sorted the blocks in descending order by the 2010 census total population, then used the logical “IF” function to examine differences between the 2010 census count and the DP count (match = zero; non-match =1), and summed the number of non-matches.

## Results from the May 27<sup>th</sup> 2020 File

### *Case 1: Children without Adults: How Did Differential Privacy turn three blocks into 765?*

The 2010 census reported that there were three blocks in which 1 or more children (under age 18) were listed, but no adults (18 years and over). Of these three blocks, the first had one child, the second, five children, and the third had 15 children. It is likely that the last block has a facility where children reside in the presence of adults who themselves live elsewhere.

Out of 45,292 blocks, it is highly believable that there are three in which a total of 21 children reside without adults. However, DP produced 765 such blocks in which 3,381 children reside without adults - a highly unbelievable number

*Case 2: Differential Privacy turned 1,252 Blocks with one or more people of voting age into blocks with zero people of voting age*

- In comparing the voting age populations reported by the 2010 census and the DP file, it was found that there are 1,252 blocks in which DP reported zero people of voting age while the 2010 census reported one or more persons of voting age in these same blocks.

*Case 3: Differential Privacy turned 830 blocks with zero persons of voting age into blocks with one or more persons of voting age*

- At the same time, DP turned 830 blocks in which the 2010 census reported zero persons of voting age into blocks with one or more persons of voting age.

*Case 4: Of 12,870 blocks in which the 2010 census shows one or more persons, 12,366 of them (96%) show a different number of persons when DP is applied.*

- Of these same 12,870 blocks, 12,009 of them (93%) show a different number of persons of voting age population (18 years and over) when DP is applied.

### **Results from the April 28<sup>th</sup>, 2021 File**

*Case 1: Children without Adults: How Did Differential Privacy turn three blocks into 428?*

The 2010 census reported that there were three blocks in which 1 or more children (under age 18) were listed, but no adults (18 years and over). Of these three blocks, the first had one child, the second, five children, and the third had 15 children. It is likely that the last block has a facility where children reside in the presence of adults who themselves live elsewhere.

Out of 45,292 blocks, it is highly believable that there are three in which a total of 21 children reside without adults. However, DP produced 428 such blocks in which 1,302 children reside without adults - a number that remains unbelievable.

*Case 2: Differential Privacy turned 533 Blocks with one or more people of voting age into blocks with zero people of voting age*

- In comparing the voting age populations reported by the 2010 census and the DP file, it was found that there are 533 blocks in which DP reported zero people of voting age while the 2010 census reported one or more persons of voting age in these same blocks.

*Case 3: Differential Privacy turned 830 blocks with zero persons of voting age into blocks with one or more persons of voting age*

- At the same time, DP turned 632 blocks in which the 2010 census reported zero persons of voting age into blocks with one or more persons of voting age.

*Case 4: Of 12,866 blocks in which the 2010 census shows one or more persons, 11,801 of them (92%) show a different number of persons when DP is applied.*



## Discussion and Conclusion

Alaska was not subject to higher levels of DP Disclosure Avoidance than the other states in either of the two “Demonstration Product” files (2020527 and 20210428) we have analyzed. Instead, the DP levels are reported as uniform across all states at an “epsilon” level of 4.0 and 10.3, respectively, for people ([https://www.nhgis.org/privacy-protected-demonstration-data#v20210428\\_12-2](https://www.nhgis.org/privacy-protected-demonstration-data#v20210428_12-2)). Given this and the low numbers of people found statewide in the 2010 census and its low number of 2010 census blocks, Alaska would appear to be a candidate for a higher level of DP Disclosure Avoidance than many other states.

Finding that in going from an epsilon of 4.0 in which DP produced 765 census blocks in which 3,381 children reside without adults to an epsilon of 10.3 in which DP produced 428 such blocks in which 1,302 children reside without adults remains very troubling, as are our other three comparisons

If DP is implemented at the avoidance level found in either of the two “Demonstration Product” files (2020527 and 20210428) for census blocks in Alaska we examined in this study, it will affect almost all of the state’s users of small area census data, from legislatures relying on the data to design Congressional Districts to comply with the law, to demographics vendors who supply clients with zip code level characteristics so businesses can make better decisions. Other end users such as health district administrators who need the data to tract health issues such as COVID-19, and businesses that use small area data such as zip codes, blocks and block groups to improve marketing stand to be dramatically impacted. Many government agencies also depend on accurate small area census data to make programs run efficiently and effectively and the biggest impact of DP will be in small areas. The data in small areas are typically used both directly where the small area is the unit of analysis and aggregated into higher levels of geography by these users. In the case of the latter, the errors introduced by DP tend to even out. However, in the case of the former, these users and their clients will be forced to deal with erroneous data if DP is implemented.

Because it is likely that the results we found in Alaska will be found in other states and perhaps at even higher levels of error, our examination leads us to conclude that it is likely the errors introduced by DP of the type and at the level found in the demonstration product file we examined will render the nation’s block level data essentially unusable.

## Acknowledgements

We are grateful to the Minnesota Population Center for assembling and making available the DP demonstration product file we use here. We also are grateful for advice and comments from Jan Vink and Bill O’Hare.

## References

Abowd, J (2020). Modernizing Disclosure Avoidance: What We’ve Learned, Where We Are Now. [https://www.census.gov/newsroom/blogs/research-matters/2020/03/modernizing\\_disclosu.html](https://www.census.gov/newsroom/blogs/research-matters/2020/03/modernizing_disclosu.html)

Beveridge, A. (2021). Controversial Census Bureau Plan That Makes Data Less Accurate Goes to Court. (<https://www.socialexplorer.com/blog/post/controversial-census-bureau-plan-that-makes-data-less-accurate-goes-to-court-11452> ).

Hunsinger, S., D. Howell, and E. Sandberg. (2012). Alaska Population Overview: 2010 Census and 2011 Estimates. Research & Analysis Section, Alaska Department of Labor, Juneau, AK (<http://live.laborstats.alaska.gov/pop/estimates/pub/1011popover.pdf> ).

Ruggles, S., C. Fitch, D. Magnuson, and J. Schroeder. (2019). Differential Privacy: Implications for Social and Economic Research. American Economic Association Papers and Proceedings 109 (May): 403-408.

U.S. Census Bureau (2020a). 2020 Census Disclosure Avoidance Improvement Metrics, U.S Census Bureau, Washington DC., March 18, <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf?#>

U.S. Census Bureau (2020b). “2020 Census Data Products and the Disclosure Avoidance System, Hawes M. and Garfinkel. S. L., Planned presentation at the Census Scientific Advisory Committee meeting, March 26,

U.S. Census Bureau (2020c). DAS Updates, U.S Census Bureau, Hawes M. June 1 Washington DC., <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-06-01-das-updates.pdf?#>

U.S. Census Bureau (2020d). “Disclosure Avoidance and the Census,” Select Topics in International Censuses, U.S. Census Bureau, October 2020. <https://www.census.gov/library/working-papers/2020/demo/disclos-avoid-census.html>

U.S. Census Bureau (2020e). “Disclosure Avoidance and the 2020 Census, U.S. Census Bureau,” Washington DC., Accessed November 2<sup>nd</sup>. [https://www.census.gov/about/policies/privacy/statistical\\_safeguards/disclosure-avoidance-2020-census.html](https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html)

U.S. Census Bureau (2020f). Error Discovered in PPM, U.S. Census Bureau, Washington DC. <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>

U.S. Census Bureau (2020g). “2020 Disclosure Avoidance System Updates,” U.S. Census Bureau, Washington DC., <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>

## **The Effect of the Differential Privacy Disclosure Avoidance System Proposed by the Census Bureau on 2020 Census Products: Four Case Studies of Census Blocks in Mississippi.**

David A. Swanson, University of California Riverside, Riverside, CA and  
The Center for Studies in Demography and Ecology, University of Washington, Seattle, WA  
(email: [dswanson@ucr.edu](mailto:dswanson@ucr.edu))

Ronald E. Cossman, Social Science Research Center, Mississippi State University, Starkville, MS  
(email: [ronald.cossman@ssrc.msstate.edu](mailto:ronald.cossman@ssrc.msstate.edu))

### **Abstract**

The Census Bureau plans to introduce a new Disclosure Avoidance System known as Differential Privacy (DP) for its 2020 census data products. Using a DP demonstration product file provided by the Census Bureau, we assess the errors introduced by DP on census block data in Mississippi in the form of four case studies and find them to be substantial by type and level. We use both the May 27<sup>th</sup> 2020 DP demonstration product and the most recent, the April 28<sup>th</sup> 2021 DP demonstration product relative to our four cases studies and compare the changes. This comparison is important because the Census Bureau reports, the accuracy should improve because the privacy budget was increased in response to user complaints about poor accuracy. We find that the April 28<sup>th</sup>, 2021 release does produce more accurate data but that the level of accuracy remains unsuitable for use by those who work with small area data. Because it is likely that the results we found in Mississippi will be found in other states, our examination leads us to conclude that it is likely that the errors introduced by DP of the type and at the level found in the most recent demonstration product file we examined will render the nation's block level data essentially unusable.

### **Introduction**

The 2020 Census attempts to count every person living in the United States and the five U.S. territories. The stated goal is to count everyone only once and in the right place. The count is mandated by the Constitution and conducted by the U.S. Census Bureau. The requirement of taking a census is one of the first things mentioned in the U.S. Constitution, which provides some indication of how important a census was to the Founding Fathers (National Research Council, 2006).

The U.S. Constitution requires an "actual enumeration" of the population every 10 years to apportion seats in the House of Representatives among the states. States and localities also use census numbers for redistricting, to draw political boundary lines for their congressional delegations, legislatures, and other government districts. The census plays an important role in guiding the distribution of \$1.5 trillion in

federal funding, as well as identifying needs for government services, such as schools and roads. Census statistics are the basis for a wide range of research and business decisions.

In a recent publication of the International Association of Official Statistics discussing the importance of Censuses in an international context, Everaers (2021) stated, “Population and Housing Censuses are an important cornerstone for National Statistical Systems. They provide a range of important statistics, relevant for policy-making, planning, and monitoring but also functioning as reference point and sample frame for many other national and regional statistics.” This description certainly applies to the U.S. Census. There is no single statistical resource more important than the Decennial Census.

In every census, the U.S. Census Bureau faces a trade-off between privacy protection and accuracy. According to the U.S. Census Bureau (2020d),

“One of the most important roles that national statistical offices (NSOs) play is to carry out a national population and housing census. In so doing, NSOs have two data stewardship mandates that can be in direct opposition. Good data stewardship involves both safeguarding the privacy of the respondents who have entrusted their information to the NSOs as well as disseminating accurate and useful census data to the public.”

The preceding suggests that this is an appropriate place to discuss privacy and confidentiality, two concepts that are often used interchangeably, but are distinct. Privacy generally is used in regard to the right of an individual or organization to withhold information from others, while confidentiality is viewed as an extension of privacy in which an organization (such as the Census Bureau) that holds individual or organizational information is obligated to ensure that only authorized individuals have access to the information.

For over a century and for nearly as long as the Census Bureau has existed in its present form, it has had to balance its inherent, ingrained mission of collecting and producing high quality statistical information for the public good with a mandate to avoid disclosing information about any individual. In fact, the

Census Bureau's mission is “to serve as the nation's leading provider of quality data about its people and economy.” However, the mandate of “quality data” is tempered by an obligation to protect the privacy of Census respondents. The Census Bureau is bound by Title XIII of the United States Code. Title XIII provides the following protections to individuals and businesses (U.S. Census Bureau, no date):

- Private information is never published. It is against the law to disclose or publish any private information that identifies an individual or business such, including names, addresses (including GPS coordinates), Social Security Numbers, and telephone numbers.
- The Census Bureau collects information to produce statistics. Personal information cannot be used against respondents by any government agency or court.
- Census Bureau employees are sworn to protect confidentiality. People sworn to uphold Title XIII are legally required to maintain the confidentiality of data. Every Census Bureau employee or contractor with access to personal data is sworn for life to protect your information and understands that the penalties for violating this law are applicable for a lifetime.

As part of this balancing act, the Census Bureau has used methods to help avoid disclosure of individual census respondents for many decades. According to the U.S. Census Bureau (2018), some method of disclosure avoidance has been used by the U.S. Census Bureau since 1970. However, as the privacy protections were put in place by the Census Bureau over the past several decades, there was never the threat of distorting the data as much as DP threatens to distort the 2020 Census data, and there was never the resistance seen among data users and demographers regarding the potential use of DP in the 2020 Census (Ruggles et al., 2019). The increase in resistance among data users reflects the extent to which they fear differential privacy will distort the data to the point that it is not usable for many functions.

The Census Bureau plans to introduce a new Disclosure Avoidance System known as Differential Privacy (DP) for its 2020 census data products (Abowd, 2020, Census Bureau 2020a, 2020b, 2020c, 2020d, 2020e, 2020f, and 2020g), which we describe in some detail in Appendix 1.

Ruggles et al. (2019: 406) argue that DP goes far beyond what is necessary to keep data safe under census law and precedent and because it focuses on concealing individual characteristics instead of respondent identities, DP is a blunt and inefficient instrument for disclosure control. They go on to note that because the core metric of DP does not measure the risk of identity disclosure, it cannot assess disclosure risk as defined under census law, making it untenable for optimizing the privacy/usability trade-off.

Our purpose in this paper is to assess the errors introduced by (DP) on census block data in Mississippi in the form of comparing four case studies taken from the May 27<sup>th</sup>, 2020 and April 28<sup>th</sup>, 2021 demonstration products, respectively.

### **Data and Methods**

Mississippi is the 32<sup>nd</sup> largest state with 48,430 square miles (<https://en.wikipedia.org/wiki/Mississippi>). The 2010 census counted 2,967,297 persons (U.S. Census Bureau, 2012: IV-3), which yields 61.3 persons per square mile. The 2010 census ~~(see below)~~, organized the state into 171,778 census blocks, of which 84,750 had one or more persons, leaving 87,028 without any population. On average, there were only 17.27 persons in each of the 171,778 census blocks.<sup>1</sup> If we look at the 84,750 census blocks with at least one person, there were 35 persons on average in each of them. These summary statistics make Mississippi one of the states in which one would expect a high level of disclosure avoidance at the block level because there are so few people on average per block. This is a point to which we return in the final section.

The application of DP is a brand-new approach for the Census Bureau and is different from all prior Census Bureau initiatives in regard to disclosure avoidance. As a component of the DP initiative, the Census Bureau has released a series of “demonstration products” (Abowd, 2020, Census Bureau 2020a, 2020b, 202c, 2020d, 2020e, 2020f, and 2020g) that allow outside analysts and stakeholders to determine for their purposes the impact DP would have on Census data. These demonstration products generally contain:

- the most common, basic demographic and housing variables

- different levels of geography
- 2010 census data as they were originally reported
- 2010 census data adjusted (perturbed) by DP

As the Census Bureau responded to User complaints about poor accuracy, the “privacy budgets” were changed in the demonstration products to provide higher levels of accuracy (Beveridge, 2021). Here, we examine the errors introduced by DP on 2010 Census block data for Mississippi in the form of four case studies. In our initial analysis, we employ the “demonstration product” for census blocks in Mississippi released May 27<sup>th</sup>, 2020, file (labeled as 2020527) with an epsilon level of 4.0 , which was downloaded from the Minnesota Population Center’s NHGIS site: <https://nhgis.org/privacy-protected-demonstration-data>. Against the results we find from the May 27<sup>th</sup>, 2020 file, we compare results from the most recent release, April 28<sup>th</sup>, 2021. (file labeled as 20210428) with an epsilon level of 10.3, which was downloaded from the same Minnesota Population Center’s NHGIS site.

In the comparative analyses for case studies 1 through 3, we employed the cross-tabulation routine found in Release 12 of the NCSST Statistical System (<https://www.ncss.com/software/ncss/> ). For case study 4, we sorted the blocks in descending order by the 2010 census total population, then used the logical “IF” function to examine differences between the 2010 census count and the DP count (match = zero; non-match =1), and summed the number of non-matches.

### **Results from the May 27<sup>th</sup> 2020 File**

#### *Case 1: Children without Adults: How Did Differential Privacy turn 10 blocks into 4,912?*

The 2010 census reported that there were 10 blocks in which one or more children (under age 18) were listed, but no adults (18 years and over). There were 371 children in these ten blocks.

Out of 171,778 blocks, it is highly believable that there are ten in which a total of 371 children reside without adults. However, DP produced 4,912 such blocks in which 27,383 children reside without adults - a highly unbelievable number

*Case 2: Differential Privacy turned 8,235 Blocks with one or more people of voting age into blocks with zero people of voting age*

*Case 3: Differential Privacy turned 1,886 blocks with zero persons of voting age into blocks with one or more persons of voting age*

*Case 4: Excluding the 84,813 blocks in which both the 2010 census and the DP Adjustment shows zero population, Mississippi has 89,966 blocks where either the 2010 census or the DP adjustment show at least one person. Of these 86,966 blocks:*

- 83,425 (96%) show a different total population when DP is applied.
- 82,821 (95%) show a different number of adults (18 years and over) when DP is applied
- 72,051 (83%) show a different number of children (under 18 years of age) when DP is applied

#### **Results from the April 28<sup>th</sup> 2021 File**

*Case 1: Children without Adults: How Did Differential Privacy turn 10 blocks into 3,100?*

The 2010 census reported that there were 10 blocks in which one or more children (under age 18) were listed, but no adults (18 years and over). There were 371 children in these ten blocks.

Out of 171,778 blocks, we repeat that it is highly believable that there are ten in which a total of 371 children reside without adults. However, the April 28<sup>th</sup> 2021 DP demonstration product yields produced 3,100 such blocks in which 9,418 children reside without adults - a highly unbelievable number

*Case 2: Differential Privacy turned 4,907 Blocks with one or more people of voting age into blocks with zero people of voting age*



*Case 3: Differential Privacy turned 3,048 blocks with zero persons of voting age into blocks with one or more persons of voting age*

*Case 4: Excluding the 84,997 blocks in which both the 2010 census and the DP Adjustment shows zero population, Mississippi has 86,801 blocks where either the 2010 census or the DP adjustment show at least one person. Of these 86,966 blocks:*

- *80,063 (92%) show a different total population when DP is applied.*
- *77,712 (90%) show a different number of adults (18 years and over) when DP is applied*
- *69,666 (80%) show a different number of children (under 18 years of age) when DP is applied*

### **Discussion and Conclusion**

As far as we can tell from the information available from the Minnesota Population Center, Mississippi was not subject to higher levels of DP Disclosure Avoidance than the other states in either of the two “Demonstration Product” files (2020527 and 20210428) we have analyzed. Instead, the DP levels are reported as uniform across all states at an “epsilon” level of 4.0 and 10.3, respectively, for people ([https://www.nhgis.org/privacy-protected-demonstration-data#v20210428\\_12-2](https://www.nhgis.org/privacy-protected-demonstration-data#v20210428_12-2)). Given this and the low numbers of people found statewide in the 2010 census and its low number of 2010 census blocks, Mississippi would appear to be a candidate for a higher level of DP Disclosure Avoidance than many other states. This makes our findings all the more worrying because they show high levels of error at the census block level even at what might be described as a low level of DP Disclosure Avoidance. Finding that in going from an epsilon of 4.0 in which DP produced 4,912 census blocks in which 27,383 children reside without adults to an epsilon of 10.3 in which DP produced 3,100 such blocks in which 9,418 children reside without adults remains very troubling, as are our other three comparisons.

As the examples show in Appendix 2, if DP is implemented at either the avoidance level found in the “Demonstration Product” files 20200527 or 20210428 for census blocks in Mississippi we examined in this study, it will affect almost all of the state’s users of small area census data, from legislatures relying on

the data to design Congressional Districts to comply with the law, to demographics vendors who supply clients with zip code level characteristics so businesses can make better decisions. Other end users such as health district administrators who need the data to track health issues such as COVID-19, and businesses that use small area data such as zip codes, blocks and block groups to improve marketing, stand to be dramatically impacted. Many government agencies also depend on accurate small area census data to make programs run efficiently and effectively and the biggest impact of DP will be in small areas. The data in small areas are typically used both directly where the small area is the unit of analysis and aggregated into higher levels of geography by these users. In the case of the latter, the errors introduced by DP tend to even out. However, in the case of the former, these users and their clients will be forced to deal with erroneous data if DP is implemented.

Because it is likely that the results we found in Mississippi will be found in other states and perhaps at even higher levels of error, our examination leads us to conclude that it is likely the errors introduced by DP of the type and at the level found in the demonstration product file we examined will render the nation's block level data essentially unusable.

#### **Appendix 1. What is Differential Privacy?**

A statement by Ben Rossi (2016) summarizes the problem with DP in regard to small areas such as census blocks: "...[I]f a database is a representative sample of an underlying population, the goal of a privacy-preserving statistical database is to enable the user to learn properties of the population as a whole, while protecting the privacy of the individuals in the sample." This statement reveals that the DP tradeoff is to make available properties of the population as a whole, while protecting the privacy of individuals. In the world of the Census Bureau, this tradeoff has been translated to mean that the population as a whole, is defined by a population at a level of geography beyond the block. The tradeoff means that a user cannot

learn properties of the population at the block level with any degree of confidence. If DP is implemented, it will affect all of the many users of small area data, to include those described earlier, the demographics vendors who supply clients with zip code level characteristics, public health and public safety organizations, and businesses that use small area data such as zip codes, school districts, and Regional Planning Organizations. The data associated with these census stakeholders are those that represent small areas directly as well as being aggregated into other small areas and into higher levels of geography. This means that DP, a statistical adjustment, will increase the error in the small area data needed by these stakeholders.

Is DP complicated? Here is a formal Definition followed by a discussion. To start, we use definition 2.4 from Dwork and Roth (2014: 17).

**Definition 2.4 (Differential Privacy).** A randomized algorithm  $\mathcal{M}$  with domain  $\mathbb{N}^{|\mathcal{X}|}$  is  $(\epsilon, \delta)$ -differentially private if for all  $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$  and for all  $x, y \in \mathbb{N}^{|\mathcal{X}|}$  such that  $\|x - y\|_1 \leq 1$ :

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

Where,

M: Randomized algorithm i.e., query (db) + noise or query (db + noise).

S: All potential output of M that could be predicted.

x: Entries in the database. (i.e., N)

y: Entries in parallel database (i.e., N-1)

$\epsilon$ : epsilon, The maximum distance between a query on database (x) and the same query on database (y).

$\delta$ : Delta, the probability of information accidentally being leaked.

This definition of DP is a measure of “How much privacy is afforded by a query?” This is an important point in that DP represents an offer of privacy according to a provable and quantifiable amount, sometimes referred to as the privacy-loss budget (Snoke and McKay, 2019). It is this probabilistic quantifiable feature that is DP’s major selling point because other forms of DAS (Disclosure Avoidance Systems) do not provide a formal quantification of the protection they offer. How does it do this? As this suggests, DP is not the

system that creates privacy; it is the system that measures privacy using the definition just given. How does DP measure privacy?

The DP algorithm gives the comparison between running a query  $M$  on database  $(x)$  and on a parallel database  $(y)$ , where the latter has one less entry than database  $(x)$ . The measure by which the full database  $(x)$  and the parallel database  $(y)$  can differ is given by Epsilon ( $\epsilon$ ) and delta ( $\delta$ ). Specifically, DP works by tying privacy to how much the answer to a question or statistic is changed given the absence or presence of the most extreme possible person in the population. This is done within a statistical framework. An example by Snoke and McKay (2019) helps to explain this. Suppose the data we want to protect is income data, and the statistic we want answered is, “What is the median income?” The most extreme person who could possibly be in any given income data could be Jeff Bezos. If he is absent or present in the data set, the median will not change much, if at all. This means that DP can provide a more accurate answer about the median income without using much privacy-loss budget.

However, what if the question is, “What is the maximum income?” Unlike the median, the answer to this question would be likely to significantly change if Bezos is absent or present in the data set. A DP algorithm would provide a less accurate answer, or require more privacy-loss budget, to answer this query and protect the extreme case, Bezos (Snoke and McKay, 2019).

So, when Epsilon ( $\epsilon$ ) is small (as shown in Definition 2.4 above), DP asserts that for all pairs of adjacent databases  $x, y$  and all outputs  $M$ , an adversary cannot distinguish which is the true database on the basis of observing the output—the probabilities are too low. That is, if we are interested in median income, it does not matter if Jeff Bezos is in or out of the data set: For this query Epsilon ( $\epsilon$ ) should be set at a high level because for a query regarding median income there is little need to “protect” the data base. This example translates formally into something like the following. When ( $\epsilon$ ) is large DP merely says that there

exists neighboring databases and an output  $M$ , for which the ratio of probabilities of observing  $M$  conditioned on the database being, respectively,  $x$  or  $y$ , is large.

However, if we interested in knowing the maximum income in the data base, it will matter if Jeff Bezos is in or out of the database. Thus, Epsilon ( $\epsilon$ ) should be set at a low value to prevent “leaking” the maximum income. However, even if Epsilon ( $\epsilon$ ) is not set low, an adversary may not have the right auxiliary information to recognize that a revealing output has occurred; or may not know enough about the database(s) to determine the value of their difference.

Thus, the DP algorithm represents a statistical adjustment in that it uses a probability framework, typically based on the Laplace probability distribution (as stated elsewhere in this report), which is used to produce the errors / noise in the data. Moreover, as noted by Ruggles et al. (2019) under DP, responses of individuals cannot be divulged even if the identity of those individuals is unknown and cannot be determined. Returning to the example of a query about maximum income, it would not matter if the identify of Bezos was not divulged; the correct answer to the question about the maximum income in a dataset would not be provided under DP.

A final important point about differential privacy is that it is applied using two different types of geography: (1) “spine” which are the core census statistical geographies such as counties, tracts, and blocks; and (2) “off-spine” which are governmental or administrative geographies such as school districts and legislative districts. The “spine” geography, particularly blocks, are important because they offer the greatest geographic granularity and are the geographies DP is actually being applied to. “Off-spine” geographies are also critically important because conceptually they could capture the best or worst pieces of statistical geography and aggregate and magnify their errors. As shown in Figure X.X (above), legislative districts, voting districts, congressional districts, places, VTDs, and ZIP codes are all “off-spine,” that is, not in the hierarchy of geographic areas for which the Top Down Algorithm (TDA) maximizes accuracy and so

are built up from the lower-level block groups and blocks. In our analysis, we only look at blocks, which is one of the spine geographies.

## **Appendix 2. Selected Examples of Small Area Census Data Users**

### **Consumer Demographics**

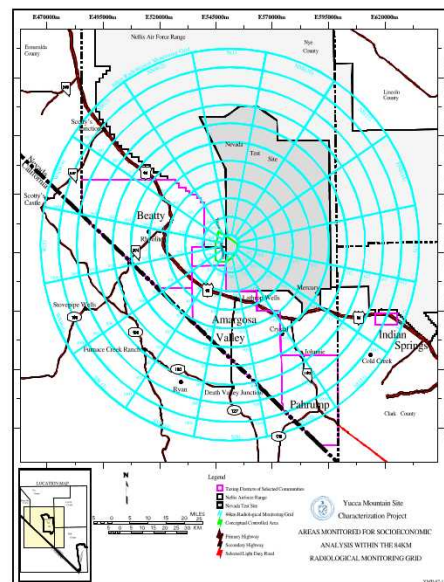
In the latter part of the 20th century, statistics became a commodity independent of government, and a statistical services industry developed (Swanson, 2013). This development is pertinent because these services are primarily a business information industry. While demographics vendors such as Claritas and ESRI generate their own zip code estimates and forecasts, the Census Bureau uses both block and block group data to generate zip code population and housing estimates (Census Bureau, 2020h). In the case of the zip code data generated by the Census Bureau, it is certain to be subject to DP if the latter is implemented; in the case of the zip code generated by demographic vendors, it is certain that the block and block group data they use in the process will be subject to DP if the latter is implemented. In 2018, for example, the Census Bureau decided to no longer approve requests for sub-state data if the data were not protected using strengthened disclosure avoidance methods providing small area data in its for-pay Custom Table operation (U.S. Census Bureau, 2020i). As an example of the importance of these data, Swanson et al. (2009) used ESRI Zip code data to assess the impacts of Hurricane Katrina in Louisiana and Mississippi.

### **Health and Safety**

Small area data are important for public health and public safety, both for planning and reporting. As one example of the use of small area data for public health, the Mississippi State Department of Health (2016) developed an integrated HIV health care and prevention plan that uses census tract data.

As another example, the U.S. Department of Energy (1988) issued a radiological monitoring plan for the investigation of a proposed nuclear waste storage at Yucca Mountain, Nevada. The radiological studies area is defined by a circle 84 km in radius, whose center is assumed to be located at the proposed site of the central surface facilities (see Figure 1 below). The circle is divided into 160 cells radiated out from the 16 km radius area in the center of the study area, designated as the near field (NF) study area. The remainder of the area (16-84 km) is called the far field (FF) study area. The FF study area required that the population of each of the 160 cells be estimated on a regular basis, which required block, block group, and census tract data as starting points (Swanson, Carlson, and Williams, 1990). This plan would have been of use in Mississippi when nuclear bombs were detonated underground in Lamar County, one in 1964 and the other in 1966 (<http://mshistorynow.mdah.state.ms.us/articles/293/nuclear-blasts-in-mississippi> ).

**Figure 1: The Radiological Studies Area.**



### **Natural Disaster Assessment**

Closely related to public health and safety, but distinct, is natural disaster preparedness and assessment. As an example, Swanson (2008) examined the effect of Hurricane Katrina on the populations of 20 selected ZIP code areas in Mississippi and found them to be profound. In another study of the demographic effects of Hurricane Katrina, Swanson (2009) examined the effects of Hurricane Katrina on the client populations and candidates for a specific medical procedure in the service areas associated with two medical facilities on the Mississippi gulf coast. The two service areas were defined by zip codes, and in analyzing them, Swanson found that Katrina had an adverse impact on the client base of both medical facilities.

As another example, what will the Census Bureau do with its Emergency Management program (<https://www.census.gov/topics/preparedness.html>), which is designed to provide timely and accurate data about the effects of natural disasters? If a Category 4 hurricane strikes Hancock, Harrison, and Jackson counties, will the Census Bureau provide erroneous small area data to FEMA and local authorities?

### **Regional Planning Organizations**

There are hundreds of regional planning organizations in the U.S. Although they exist in every state, they may come under different names in different states, (Council of Government (COG), Metropolitan Planning Organizations (MPO)) but they all have similar missions, centered on land use and transportation planning, both of which require small area data.

In conjunction with the Gulf Coast Planning Organization, the Kirwan Institute developed an index by which the geography of opportunity in the Mississippi Gulf Coast region can be viewed using census tract data (Kirwan Institute, 2012). The Central Mississippi Planning and Development District (no date) reports income data for block groups and persons per square mile by census tract. Faulty block-level data would result in the misdirection of resources intended for those populations in greatest need.



## Acknowledgements

We are grateful to the Minnesota Population Center for assembling and making available the DP demonstration product file we use here.

## Endnotes

1. Census blocks are statistical areas bounded by visible features such as roads, streams, and railroad tracks, and by nonvisible boundaries such as property lines, city, township, school district, county limits and short line-of-sight extensions of roads.

The building blocks for all geographic boundaries the Census Bureau tabulates data for, such as tracts, places, and American Indian Reservations.

Generally small in area. In a city, a census block looks like a city block bounded on all sides by streets. Census blocks in suburban and rural areas may be large, irregular, and bounded by a variety of features, such as roads, streams, and transmission lines. In remote areas, census blocks may encompass hundreds of square miles.

## References

Abowd, J (2020). Modernizing Disclosure Avoidance: What We've Learned, Where We Are Now. [https://www.census.gov/newsroom/blogs/research-matters/2020/03/modernizing\\_disclosu.html](https://www.census.gov/newsroom/blogs/research-matters/2020/03/modernizing_disclosu.html)

Central Mississippi Planning and Development District (no date). Demographic Data (<http://cmpdd.org/demographic-data/>)

Dwork, C. and A. Roth. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science 9 (3-4): 211-407 (<https://www.cis.upenn.edu/~aaroht/Papers/privacybook.pdf>)

Everaers, P. (2021) Editorial, Statistical Journal of the International Association of Official Statistics: Statistical Journal of the IAOS, vol. 36, no. 1, pp. 1-3, 2020 DOI-10.3233/SJI-209002.

Kirwan Institute (2012). The Geography of Opportunity in Mississippi's Gulf Coast Region. Columbus, OH. ([https://grpc.com/wp-content/uploads/2018/04/FinalReport\\_Dec2012-2reduced.pdf](https://grpc.com/wp-content/uploads/2018/04/FinalReport_Dec2012-2reduced.pdf)).

Mississippi State Department of Health. (2016). 2017-2021 Integrated HIV Prevention and Care Plan (<https://msdh.ms.gov/msdhsite/static/resources/7022.pdf> )

National Research Council (2006). Once, Only Once, and in the Right Place: Residence Rules in the Decennial Census. Panel on Residence Rules in the Decennial Census. Daniel L. Cork and Paul R. Voss, eds. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Ruggles, S., C. Fitch, D. Magnuson, and J. Schroeder. (2019). Differential Privacy: Implications for Social and Economic Research. American Economic Association Papers and Proceedings 109 (May): 403-408.

Snoke, J., and C. McKay. (2019). Differential Privacy: What is it? AMSTAT News (March). (<https://magazine.amstat.org/blog/2019/03/01/differentialprivacy/> ).

Swanson, D.A. (2013). Consumer Demographics: Welcome to the Dark Side of Statistics. Radical Statistics 108: 38-46.

Swanson, D. A. (2009). "Hurricane Katrina: A Case Study of Its Impacts on Medical Service Providers and Their Client Populations." The Open Demography Journal 2: 8-17.

Swanson, D. A. (2008) "The Demographic Effects of Hurricane Katrina on the Mississippi Gulf Coast: An Analysis by Zip code." Journal of the Mississippi Academy of Sciences. 53 (4): 213-231.

Swanson, D., J. Carson, and C. Williams. (1990). "The Development of Small Area Socioeconomic Data to be Utilized for Impact Analysis: Rural Southern Nevada." Pp:985-990 in High Level Radioactive Waste Management: Proceedings of the 1990 International Conference, American Nuclear Society and American Society of Civil Engineers, New York, New York, 1990.

Swanson, D., R. Forgette, J. McKibben, M. Van Boening, and L. Wombold. (2009). "The Socio-Demographic and Environmental Effects of Katrina: An Impact Analysis Perspective". The Open Demography Journal. 2 (11): 36-46.

U.S. Census Bureau (no date), Title 13, U.S. Code ([https://www.census.gov/history/www/reference/privacy\\_confidentiality/title\\_13\\_us\\_code.html](https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html)).

U.S. Census Bureau (2012.) 2010 Census of Population and Housing, Population and Housing Unit Counts, CPH-2-26, Mississippi. U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2018), "Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing," THE RESEARCH AND METHODOLOGY DIRECTORATE, McKenna, L. U.S. Census Bureau, Washington DC., <https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf>

U.S. Census Bureau (2020a). 2020 Census Disclosure Avoidance Improvement Metrics, U.S Census Bureau, Washington DC., March 18, <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf?#>

U.S. Census Bureau (2020b). “2020 Census Data Products and the Disclosure Avoidance System, Hawes M. and Garfinkel. S. L., Planned presentation at the Census Scientific Advisory Committee meeting, March 26.

U.S. Census Bureau (2020c). DAS Updates, U.S. Census Bureau, Hawes M. June 1 Washington DC., <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-06-01-das-updates.pdf?#>

U.S. Census Bureau (2020d). “Disclosure Avoidance and the Census,” Select Topics in International Censuses, U.S. Census Bureau, October 2020. <https://www.census.gov/library/working-papers/2020/demo/disclos-avoid-census.html>

U.S. Census Bureau (2020e). “Disclosure Avoidance and the 2020 Census, U.S. Census Bureau,” Washington DC., Accessed November 2<sup>nd</sup>. [https://www.census.gov/about/policies/privacy/statistical\\_safeguards/disclosure-avoidance-2020-census.html](https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html)).

U.S. Census Bureau (2020f). Error Discovered in PPM, U.S. Census Bureau, Washington DC. <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>).

U.S. Census Bureau (2020g). “2020 Disclosure Avoidance System Updates,” U.S. Census Bureau, Washington DC., <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-censusdata-products/2020-das-updates.html>).

U.S. Census Bureau (2020h). Zip Code Tabulation Areas, ZCTAs <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html> ).

U.S. Census Bureau (2020i). Custom Tables <https://www.census.gov/programs-surveys/acs/data/custom-tables.html>).

U.S. Department of Energy (1988). NNWSI Project Environmental Field Activity Plan for Radiological Studies <https://www.nrc.gov/docs/ML0037/ML003753101.pdf> ).



# POLIDATA® Political Data Analysis

DATABASE DEVELOPMENT, ANALYSIS AND PUBLICATION;  
POLITICAL AND CENSUS DATA; LITIGATION SUPPORT

---

CLARK BENSEN

---

POLIDATA LLC • 1303 HAYWARD RD, P.O. BOX 530 • CORINTH, VT 05039

Tel: 703-690-4066 • Fax: 202-318-0793 • email: [clark@polidata.org](mailto:clark@polidata.org)

PUBLISHER OF THE POLIDATA® DEMOGRAPHIC AND POLITICAL GUIDES AND ATLASES

Honorable Steven Dillingham, Director  
U. S. Bureau of the Census  
4600 Silver Hill Road  
Washington, DC 20233

10 Apr 2020

Re: DAP2020

Director Dillingham,

This letter raises some concerns that I, as one who has been involved in districting projects since the 1980 Census, has about the Disclosure Avoidance Program (DAP). This is briefly described on a Bureau webpage entitled, "Statistical Safeguards":

*Before we publish any statistic, we apply safeguards that help prevent someone from being able to trace that statistic back to a specific respondent.*

*We call these safeguards "disclosure avoidance," although these methods are also known as "statistical disclosure controls" or "statistical disclosure limitations."*

*Although it might appear that a published table shows information about a specific individual, the Census Bureau has taken steps to disguise the original data in such a way that the results are still useful. These steps include using statistical methods such as "data swapping" and "noise injection."*

Before Census 2000 a similar issue faced the Bureau with regards to adjustment of the census counts. Congress enacted a statute<sup>1</sup> which addressed "Statistical Sampling or Adjustment" in the decennial. Important concerns of Congress expressed in the findings to PL105-119 are: "(5) the decennial enumeration of the population is one of the most critical constitutional functions our Federal Government performs; (6) it is essential that the decennial enumeration of the population be as accurate as possible, consistent with the Constitution and laws of the United States[;]."

The Supreme Court addressed that situation in an opinion announced on January 25, 1999<sup>2</sup>, 14 months before Census Day 2000, "States use the population numbers generated by the federal decennial census for federal congressional redistricting. See *Karcher v. Daggett*, 462 U. S. 725, 738 (1983) ("[B]ecause the census count represents the 'best population data available,' . . . it is the only basis for good-faith attempts to achieve population equality" . . .).

While the Commerce case focused largely on sampling, the act is more expansive and another of its findings is: "(7) the use of statistical sampling or statistical adjustment in conjunction with an actual enumeration to carry out the census with respect to any segment of the population poses the risk of an inaccurate, invalid, and unconstitutional census[;]."

A review of the language in section (h) of the findings provides a definition of what the term 'statistical method' means. This definition includes "or any other statistical

---

<sup>1</sup> See Pub. L. 105-119; Sec. 209 (a) (5) [congressional findings] Statistical sampling or adjustment in decennial enumeration of population; <https://uscode.house.gov/statviewer.htm?volume=111&page=2480>

<sup>2</sup> See *Department of Commerce v. United States House of Representatives*, 525 US 316 (1999); (98-404); argued November 30, 1998; decided January 25, 1999.

procedure, including statistical adjustment, to add or subtract counts to or from the enumeration of the population as a result of statistical inference[.]”

My main concern is with respect to districting<sup>3</sup> and is that if the Bureau implements the DAP as it is currently envisioned the thousands of entities across the nation that are responsible for revising current, or creating new district, boundaries for representative government at the state and local level will not have the “best population data available” and therefore will not be able to make good-faith attempts towards equality. I offer these comments with the understanding that many of the general concerns will be shared by numerous redistricting stakeholders once they know about DAP. Moreover, I believe there is general agreement regardless of political affiliation on this issue.

This is simply a question of process. The entities responsible for districting need to know, before the numbers are released in less than one year, that the numbers they receive will be sufficient to meet their critical need and that their own election calendars will not be disrupted by additional litigation over the numbers used to distribute political representation across their states or localities.

This is not a concern about the goals of the DAP to avoid inadvertent disclosure of personally identifiable information (PII). I believe there is substantial agreement that the privacy of certain individuals is a laudable aim in 2020<sup>4</sup>. However, it appears that the DAP presents a fundamental interference with the constitutional purposes of apportionment by reliance upon a statutory concern relating to privacy.

While a supplement to this letter will discuss some of the concerns shared by redistricting stakeholders, they will be listed below.

- 1) Adjusted numbers will not be “the best available population data”.
- 2) Stakeholders will be unable to “make good faith efforts” at equality.
- 3) Use of such a statistical method “poses the risk of an inaccurate, invalid, and unconstitutional census”.
- 4) Additional litigation over the numbers may result in distraction, delay, and costs to many districting entities.
- 5) The confidence amongst state and local governmental entities in the entire census process may be severely undermined.
- 6) While the Bureau is a national statistical agency, first and foremost it is the compiler of the “actual Enumeration” to fulfill the constitutional mandate.
- 7) Previous methods for disclosure avoidance were less pervasive. Because the previous methods were simpler techniques such as data swapping, rounding, top-coding, etc., the degree to which information was adjusted for protection was much less. On the other hand, the DAP for 2020 will affect every level of geography and the population counts.
- 8) Relative inaccuracy and bias in the DAP: “The new method allows us to precisely control the amount of uncertainty that we add according to privacy requirements.”

As discussed above, the implementation of the DAP is quite likely to affect redistricting stakeholders across the nation. It appears that there are several options available to the Bureau at this point.

- 1) **Continue with research but still implement DAP.** Of course, the Bureau could discount the concerns of the (currently) small group of stakeholders and local statistical entities and

---

<sup>3</sup> However, given the feedback from the so-called Demonstration Data during 2019 there are other concerns, such as distribution of intergovernmental aid, that may motivate others to comment on the DAP.

<sup>4</sup> Nevertheless, privacy was not an issue when the census was first taken. In fact, the first Census Act required the schedules to be posted for public review before they were submitted to the federal marshal. Specific requirements for privacy appear to have first been codified for the 1880 Census.

proceed as currently planned. Nevertheless, based upon the most recent information from working groups it appears that while improvements may be made to the range of error introduced by noise injection, the counts will still not be available for most levels of geography.

- 2) **The Black Box Engine.** Some observers have suggested that districting entities could submit any plan of interest to a website whereby the unadjusted counts could be applied and thus the plan drafters could know expeditiously how far off their numbers were from equality. Aside from the obvious logistical issues for such a process it fails for the want of transparency.
- 3) **Reduce the cross tabulations of data tables.** This could apply in a general sense to whatever cross tabulations that the Bureau provides. Such breakdowns appear to be largely developed by the Bureau for the use of federal, state, and local governments in their mission to fulfill their requirements for purposes other than apportionment.
- 4) **Reduce the breakdowns of data tables into fewer cells.** The critical dataset for redistricting, the so-called PL94 dataset<sup>5</sup> was, prior to Census 2000, a fairly simple dataset with a much smaller set of variables. With the addition of the multi-race response options in 2000 the number of data cells for the PL dataset expanded greatly. On its face this presents numerous privacy concerns even for areas that have a substantial number of persons because all six races are tallied for all multiple combinations. The level of detail in the PL94 dataset for each record is not needed by most districting entities and could be collapsed substantially and then DAP adjustments as previously done to the characteristic data could be undertaken.
- 5) **Invariant Block Counts without Characteristic Information.** Another alternative would be to hold invariant the counts of population and housing<sup>6</sup> and to simply provide no characteristic information at the block level. Choices for such an alternative could be a) include characteristic data only for areas at a specified geographic level or with counts above a threshold, as has been done with Special Tabulations previously, and/or b) have districting entities rely upon characteristic data from the American Community Survey (ACS).

Clearly, the perspective of districting stakeholders and local planning agencies is likely to something other than Option 1<sup>7</sup>. Because districting is done for so many types of entities there are varying degrees of resources and needs. Yet, considering the range of variations that are likely to be seen when a user compares the adjusted numbers to information they have independently collected over the decade, there are going to be a lot of queries. One would expect that local officials may find significant differences because they can spend the time to review the information, block by block. What does the Bureau propose for the Count Question Resolution process for Census 2020?

Other stakeholders may weigh in on this issue as well offering different options or perspectives. However, Options 4 and 5 at least appear to several stakeholders as being viable options. Option 4 could impose a burden on a relatively small number of entities but may not appease the concerns of the Bureau for privacy. Option 5 would affect substantially more entities but at least there is some alternative source of data that would provide less precision for the characteristic data and more statistical analysis for districting entities to comply with Voting Rights Act concerns. Nevertheless, even accepting Options 4 or 5 would be a substantial compromise for some stakeholders but if the only viable option for privacy is the DAP many stakeholders would likely choose one of the above or some other alternative not yet discussed.

---

<sup>5</sup> See Pub. L. 94-171. <https://uscode.house.gov/statutes/pl/94/171.pdf>

<sup>6</sup> Total Population and Voting Age Population, as well as the information on Housing Units and Group Quarters.

<sup>7</sup> N.b., while there may not be much difference of opinion about the overall concern, there may well be with respect to options.

*Concerns about Disclosure Avoidance Program*  
*Polidata ® Political Data Analysis, Clark H. Bensen, Page 4*

Respectfully yours,

*/s/ Clark H. Bensen*

Clark H. Bensen

Enclosures:

1) Supplement

[2020-0410a]

CC:

Honorable Wilbur Ross, Secretary  
U.S. Department of Commerce  
1401 Constitution Ave, NW  
Washington, DC 20230

## SUPPLEMENT

**Introduction.** For the sake of readers of this letter for whom Disclosure Avoidance is a new concept the following brief summary is provided. It is important to understand the widespread degree to which the counts from the ‘actual Enumeration’ are likely to be affected by the DAP.

In December of 2019 a conference was held that reviewed the results from the Bureau’s efforts of the application of the DAS to the 2010 Census data. Based upon information published by the Bureau during October 2019<sup>8</sup> and additional material published subsequent to the December 2019 conference and recent meetings of the Expert Group (which now includes at least one for redistricting) it is still unclear exactly what the actual plan for the Bureau is or will be. Moreover, it appears that the current schedule is that final policy decisions will not be made, for the design of the DAS, until September 2020<sup>9</sup>.

Currently, the best information of the degree to which numbers eventually reported for the 2020 Census can only be gleaned from the information provided in the October 2019 memo which detailed the status of these numbers for the review of the 2010 Census data. In other words, the plan, at that point, was that some numbers would be ‘invariant’, that is, the reported number would be the enumeration counts and no alteration for privacy would be made, while others will be ‘variant’, that is, the numbers reported would be altered for privacy protection.

That proposal would treat only three types of counts as invariant: a) the state total population; b) the number of housing units in a census block; and c) the number and type of group quarters in a census block<sup>10</sup>. In other words, below the state, every number provided by the Bureau will not be a tabulation of the responses from an ‘actual Enumeration’ but the result of a statistical alteration. “Differential privacy allows us to inject a precisely calibrated amount of noise into the data to control the privacy risk of any calculation or statistic.”<sup>11</sup>

Additionally, there is the question as to which metrics will be released with the adjusted numbers to allow users to assess the degree to which noise has been added. A recent March 2020 presentation<sup>12</sup> primarily addressed “making population counts more accurate” and reviewed numerous metrics that might “allow the public to see the improvements that are made” as the Bureau continues to test their DAS operations.

At this point it is an open question as to whether this will substantially change so that the block counts would be delivered as enumerated or adjusted. Regardless, what this indicates is that we are now less than one year away from releases of the numbers and the Bureau still does not know with any precision what method they will use or metrics they will provide. Notably, the implementation of disclosure avoidance will not be applied to the American Community Survey (ACS) until 2025<sup>13</sup>. Why is it that the purposes of apportionment will be the first real test case for such a statistical adjustment?

---

<sup>8</sup> See Memorandum 2019.25: 2010 Demonstration Data Products – Design Parameters and Global Privacy-Loss Budget; [https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019\\_25.html](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_25.html)

<sup>9</sup> See Updates and DAS Development Schedule, March 18, 2020; <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-updates-das-development-schedule.pdf?#>

<sup>10</sup> See the Bureau site: [https://www.census.gov/about/policies/privacy/statistical\\_safeguards/disclosure-avoidance-2020-census.html](https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html)

<sup>11</sup> See the Bureau site: <https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf?#>

<sup>12</sup> See 2020 Census Disclosure Avoidance Improvement Metrics; <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf?#>

<sup>13</sup> See <https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html>



One issue that appears to have concerned the Bureau over the threat of what they term as reconstruction of the census appears to be the result of the extraordinary level of detail that is provided by two data products: a) the block-level data provided pursuant to PL94-171 and b) the numerous cross-tabulation tables that are provided by the Bureau of data at numerous levels of census geography.

The block-level data is the critical dataset for most redistricting stakeholders. Blocks have a huge range of population across any geographic area. Many have no population because they are industrial areas, or parks, or bodies of water, or highways, or mountains, or wide open range, or simply vacant housing. Many have a handful, and many have thousands, of persons. But, blocks are used as the lowest level for most districting datasets, generally because of a few factors that make them unusual amongst all the so-called 'summary levels' that the Bureau recognizes.

These characteristics of census blocks include:

- 1) they are the lowest level for which the counts have heretofore been tabulated and made available;
- 2) they cover the entire non-coastal geographic area of a state or locality;
- 3) pursuant to the Block Boundary Suggestion Project (BBSP) the states have the ability to designate the boundaries of the blocks;
- 4) tabulations generally account for how the block fits into higher levels of geography, such as Voting Districts (VTDs) the boundaries of which are designated by many states as Phase 2 of the BBSP;
- 5) the reported counts for every higher level of geography has been simply the sum of the information for all corresponding blocks;
- 6) redistricting stakeholders form one of the few groups that rely upon the block-level information as the critical data needed to fulfill their need, that is, the purposes of apportionment; equalizing population would be considerably more difficult if higher level information was the only level for which accurate data were available<sup>14</sup>.

Below are some notes on the concerns enumerated in the letter.

- 1) Adjusted numbers will not be "the best available population data".
  - a. This is the language used in the *Karcher* case which was quoted by the SCOTUS in the Commerce Department opinion in 1999 about adjustment.
  - b. The basic concern here is that the both phases of the apportionment process, i.e., the apportionment of seats to predetermined units (e.g., states) and the districting phase should rely upon the best available data.
  - c. The Bureau has indicated that the state-level counts would be held invariant; a position that changed after initial discussions with stakeholders.
- 2) Stakeholders will not be able to "make good faith efforts" at equality.
  - a. This language also refers to the *Karcher* case which basically requires a zero-tolerance for population amongst congressional districts.
  - b. Also of note are the *Larios v. Cox* case (out of Georgia) in 2004<sup>15</sup> and the *Tennant v. Jefferson County Commission* case (out of West Virginia) in 2012<sup>16</sup>. Larios reiterated the focus of the reapportionment cases of the 1960s that the goal (therein for legislative districting) was to have equally populated districts.

---

<sup>14</sup> Note also that blocks are numbered by the Bureau and thus Block Groups, the next higher level above Blocks, are simply agglomerations of adjacent Blocks for statistical purposes. Census Tracts, the next level up the main hierarchy (aka the Spine) are designed to be generally consistent over time but have, on average thousands of persons.

<sup>15</sup> See *Cox v. Larios*, 542 US 947 (2004); no. 03-1413, decided June 30, 2004;

<sup>16</sup> See *Tennant v. Jefferson County Commission*, 567 US 758 (2012); no. 11-1184; decided September 25, 2012.

- c. The West Virginia case muddied this up a bit (for congressional districting) allowing some leniency for population deviation based upon the competing interests of the lowest deviation and legitimate state objectives. In reality this opinion reminded stakeholders of the original perspective of the Court in *Karcher*.
- 3) Use of such a statistical method “poses the risk of an inaccurate, invalid, and unconstitutional census”.
  - a. In its findings, the Congress was apparently referring to the competing analyses of the proposed adjustment for undercount which adjustment was to be based upon a statistical method known as sampling.
  - b. The Commerce case hinged largely on the statutory interpretation of the Census Act in sections 141 and 195 and held that the statistical method known as sampling was not an available method for the numbers compiled for the purposes of apportionment.
- 4) Additional litigation over the numbers may result in distraction, delay, and costs to many districting entities.
  - a. National entities are frequently at the forefront of litigation over these types of issues and bear the cost of having the courts reach a generally applicable ruling. However, given the range of error that might be infused into the process by noise injection it is likely that numerous cases may occur because of a dispute over how to interpret the altered numbers. The burden and confusion in such cases may redound to localities that may not be able to afford litigation through the entire process.
- 5) The confidence amongst state and local governmental entities in the entire census process may be severely undermined.
  - a. Local officials will review the census results block-by-block and when they discover that the reported results are different, and frequently substantially so, they will be concerned.
  - b. In recent censuses there has been a Count Question Resolution Program (CQR) to review the counts upon request and correct them if and as needed. It is unclear how this can be implemented if DAP is used for 2020.
- 6) While the Bureau is a national statistical agency, first and foremost it is the compiler of the “actual Enumeration” to fulfill the constitutional mandate.
  - a. There appears to be a break in the internal firewall at the Bureau vis-à-vis fulfillment of the constitutional mandate and ongoing survey programs. Admittedly, the number of survey programs that are done for other agencies and those that present the demographics of the nation to the world are the everyday projects for much of the Bureau. Understandably, what is good enough for a statistical agency to present may fall short of the standard of care for the counts used for “the purposes of apportionment”.
  - b. Of course, there are some projects that focus on the high quality of the actual enumeration at the Bureau and Complete Count Committees, as well as NGOs, work diligently throughout the decade to make the decennial “the best population data available”. Implementing DAP may lessen that focus because the numbers that will be used for redistricting will not be from the enumeration but altered in the manner proposed by the data scientists and decided by the Disclosure Review Board.
- 7) Previous methods for disclosure avoidance were less pervasive.
  - a. The previous methods were simpler, and easily understandable, techniques such as data swapping, rounding, top-coding, etc. and the degree to which all census

*Concerns about Disclosure Avoidance Program*  
*Polidata ® Political Data Analysis, Clark H. Bensen, Page 8*

information was adjusted for protection was much less. On the other hand, the DAP for 2020 will affect almost every level of geography and the population counts.

- b. The DAP really is a 'sea change' for redistricting and the census. Users of the special tabulations have accepted previous efforts at disclosure avoidance because those users are cognizant of the problems and the shortcomings in protected data for their specific purpose, which would rarely require the precision needed for the purposes of apportionment.
- 8) Relative inaccuracy and bias in the DAP.
- a. "The new method allows us to precisely control the amount of uncertainty that we add according to privacy requirements." Not only will the data scientists determine the best method to adjust the counts but there will inevitably be some loss of accuracy which will have some level of bias for or against some subgroup of the census universe.
  - b. It is still unclear exactly what this bias will be at this point but what is likely is that once a bias is anticipated or observed the question of using the DAP will no longer be simply one of process but a political fight of the disfavored groups against the favored groups.

###

[2020-0410a]

## COLORADO GENERAL ASSEMBLY

EXECUTIVE COMMITTEE  
Rep. KC Becker, Chair  
Sen. Leroy Garcia, Vice Chair  
Sen. Stephen Fenberg  
Rep. Alec Garnett  
Sen. Chris Holbert  
Rep. Patrick Neville

STAFF  
Natalie Mullis, Director



## EXECUTIVE COMMITTEE OF THE LEGISLATIVE COUNCIL

ROOM 029 STATE CAPITOL  
DENVER, COLORADO 80203-1784  
E-mail: [ics.ga@state.co.us](mailto:ics.ga@state.co.us)  
303-866-3521 FAX: 303-866-3855

June 1, 2020

Dr. Steven Dillingham  
Director, U.S. Census Bureau  
4600 Silver Hill Road  
Washington, D.C. 20233

Dr. Dillingham,

We thank you for the opportunity to comment on differential privacy and for providing the 2010 Demonstration Data Products. The Census Bureau provides critical information to states. We greatly appreciate the services you provide and your efforts to seek new ways to protect the privacy of survey respondents.

Colorado's State Demographer's Office has analyzed the 2010 Demonstration Data Products and presented [their findings](#) to us.<sup>1</sup> We have identified the following patterns based on this analysis:

- A general shift in population from urban areas to rural areas;
- Inaccurate population counts at sub-state levels, and a mismatch between population data and housing-level data;
- Nonsensical population placements (e.g., persons in Census Blocks where housing units are not present); and
- Significant distortions in sub-state demographic data, including statistics on age, sex, and race.

---

<sup>1</sup> A Colorado map of differences in population totals and other comparisons between the 2010 Census and demonstration file for counties, census places, legislative districts, census tracts and census blocks is available at <http://arcg.is/1X4afz>.

The analysis supporting these findings are documented in the attached materials prepared by Colorado's State Demographer's Office, including a one-page "fact sheet" and selected maps (Figures 1 through 4).<sup>1</sup>

The data distortions found in the attached analysis pose the following consequences for the State of Colorado:

- Legal implications, including violation of the United States and Colorado Constitutions, for the state legislative and congressional redistricting process because population, race, and ethnicity counts are not accurate;
- Changes in population-based distributional formulas for federal and state grants and other funding allocations to local governments that are not proportional to actual population counts;
- Inaccurate analyses used to inform public policymaking due to distortions in data, including economic, demographic, household and public health data;
- Reduced confidence in Census Bureau and other government data; and
- Less informed policymaking without a reliable alternative to Census and Census-dependent data.

Given these findings and their consequences for Colorado, we strongly recommend that the Census Bureau hold population and household data invariant at the census block level, and pursue efforts to maintain the accuracy of demographic data (age, race/ethnicity, and sex) at the sub-state level, including counties and census places.

Sincerely,



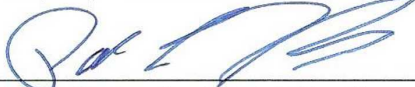
---

Speaker K.C. Becker, Chair



---

Majority Leader Alec Garnett



---

Minority Leader Patrick Neville



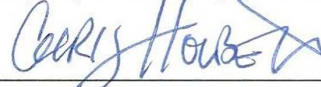
---

President Leroy Garcia, Vice-Chair



---

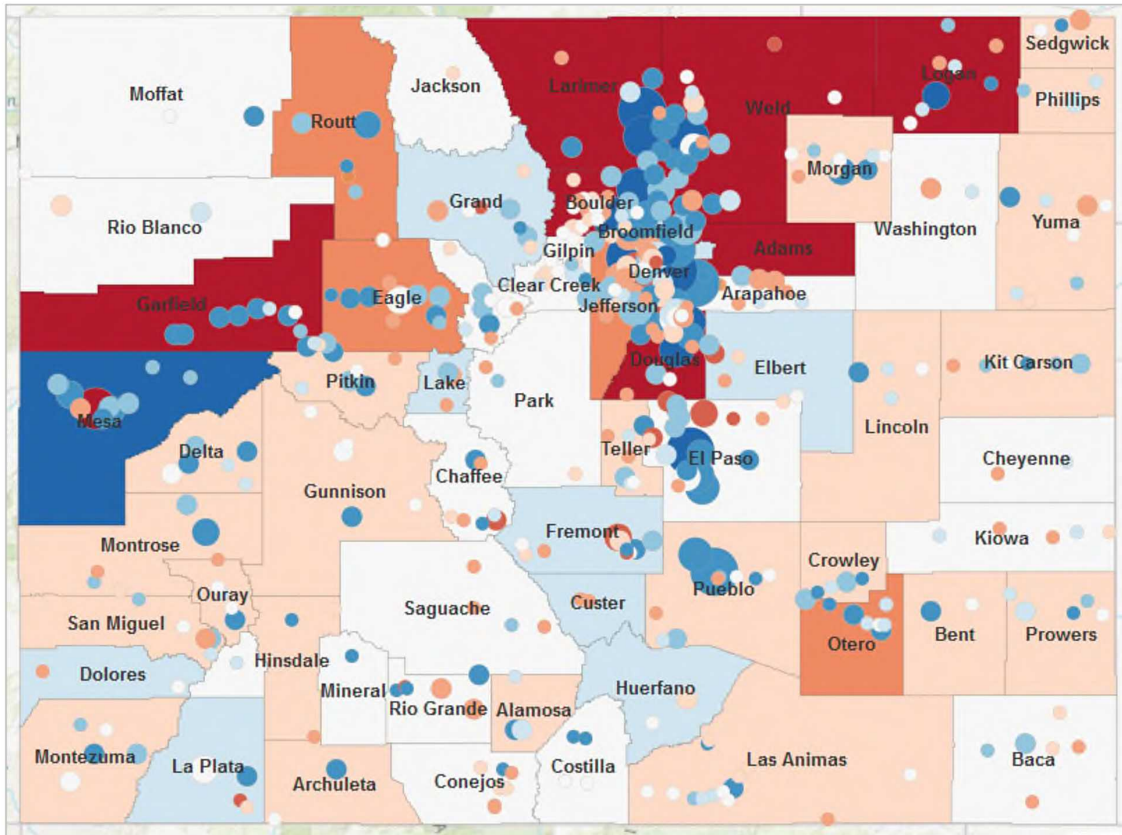
Majority Leader Stephen Fenberg



---

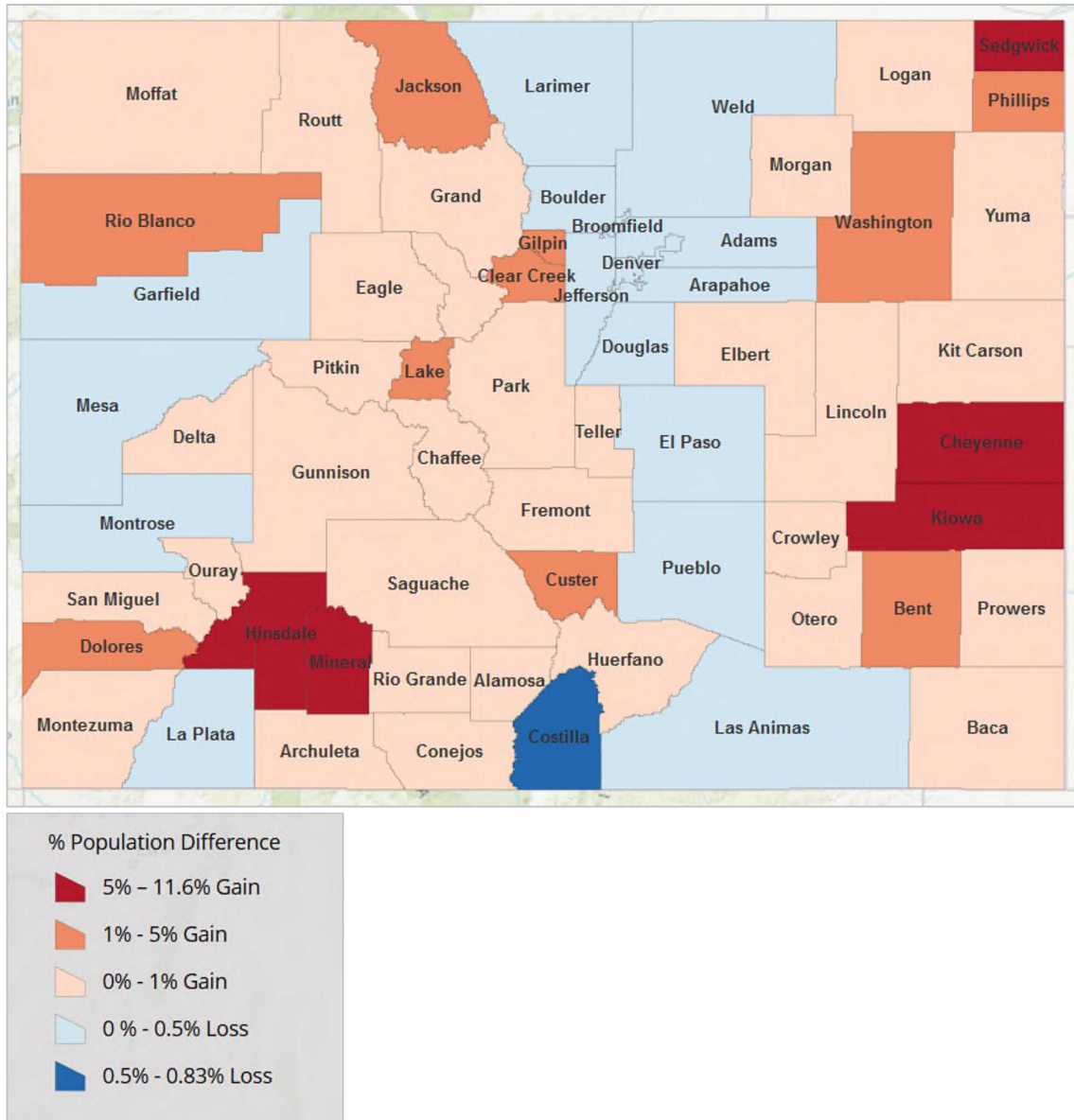
Minority Leader Chris Holbert

**Figure 1**  
**Population Differences across Census Places\***  
*Places (Bubbles) and Unincorporated Areas (County Boundaries)*



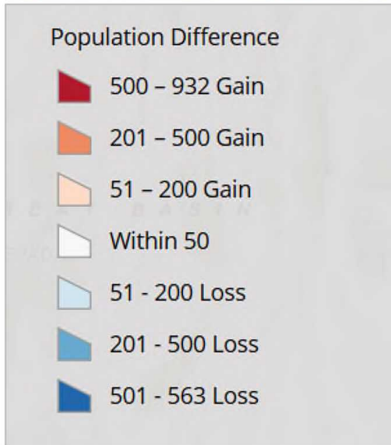
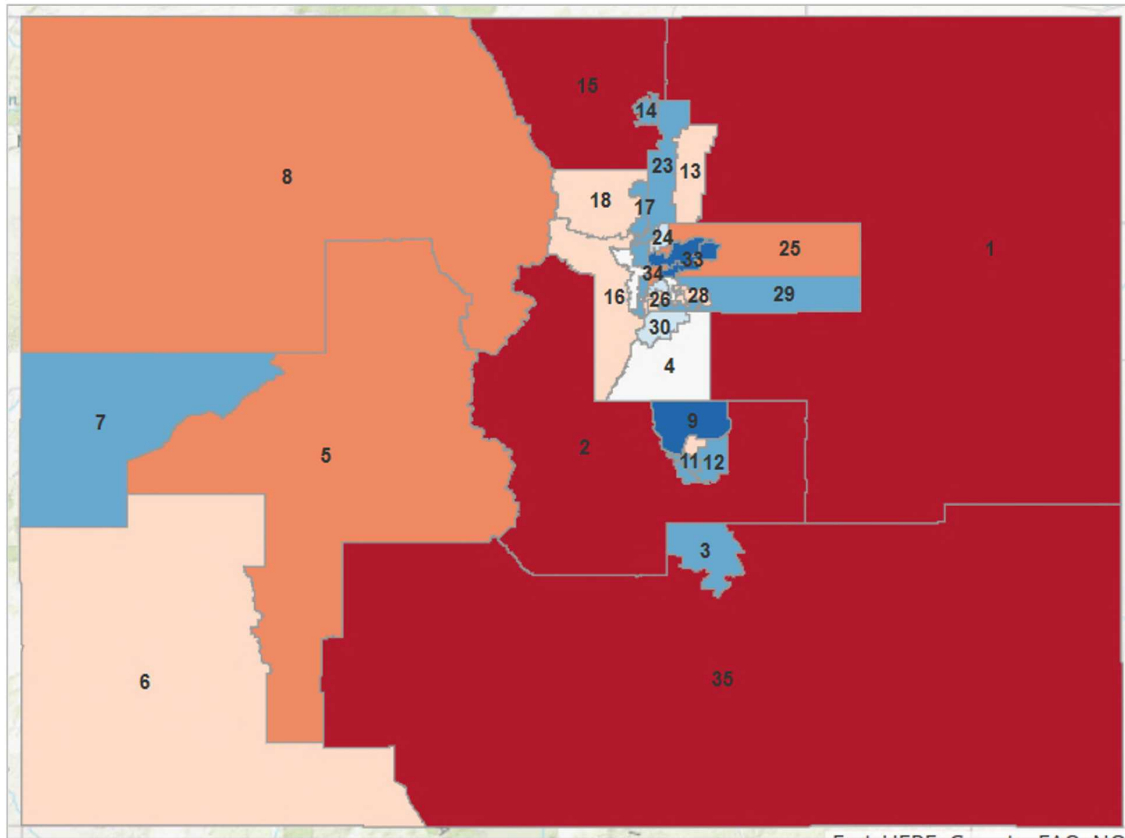
Source: Colorado Department of Local Affairs, State Demography Office. Available at: <http://arcg.is/1X4afz>  
 \*Difference in population totals between the 2010 Census and differential privacy demonstration file.

**Figure 2**  
**Population Differences across Counties\***



Source: Colorado Department of Local Affairs, State Demography Office. Available at: <http://arcg.is/1X4afz>  
 \*Difference in population totals between the 2010 Census and differential privacy demonstration file.

**Figure 3**  
**Population Differences across Colorado Senate Districts\***

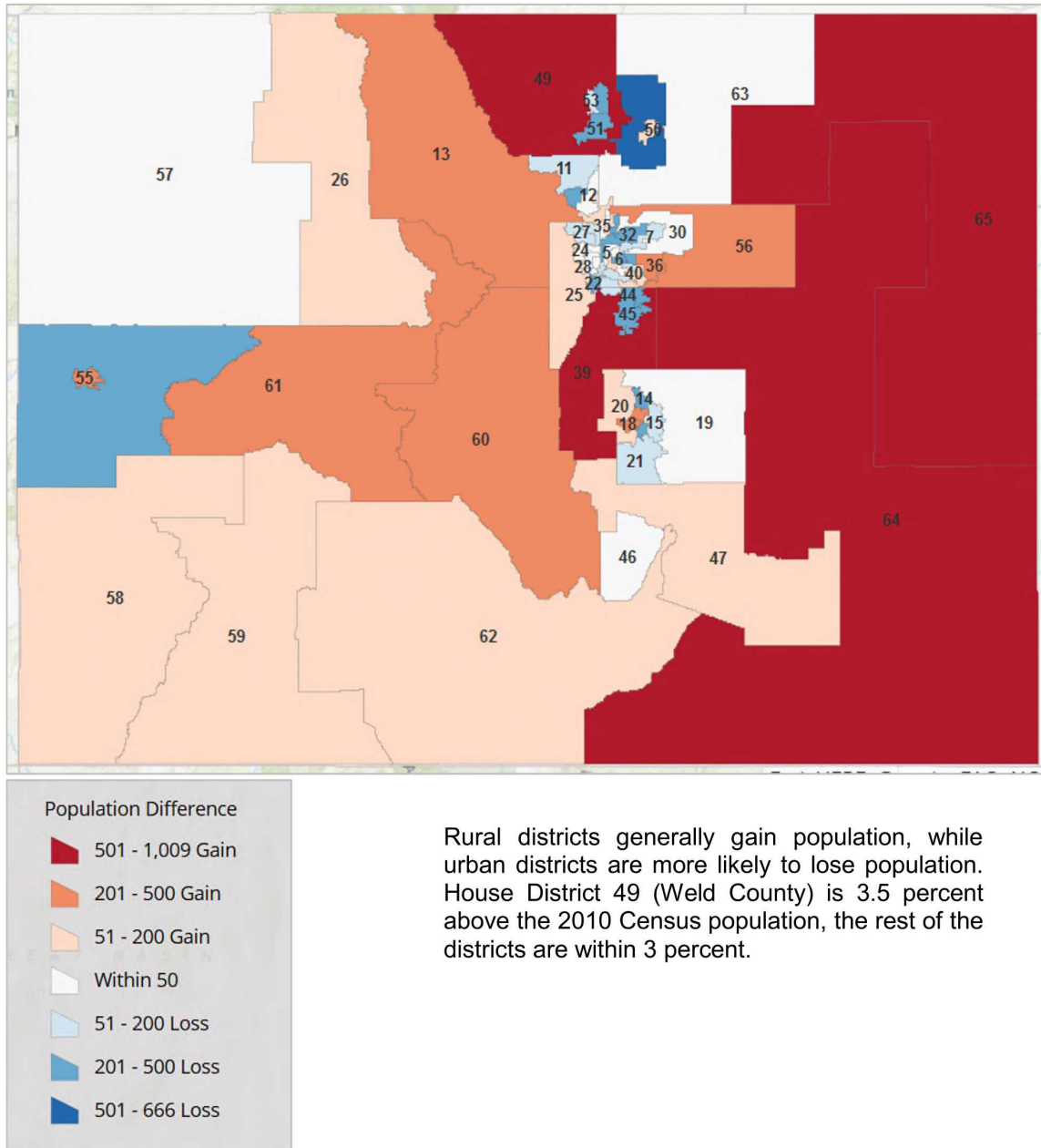


Rural districts generally gain population, while urban districts are more likely to lose population. The largest variation from the 2010 Census population is 2.8 percent.

Source: Colorado Department of Local Affairs, State Demography Office. Available at: <http://arcg.is/1X4afz>  
 \*Difference in population totals between the 2010 Census and differential privacy demonstration file.



**Figure 4**  
**Population Differences across Colorado House Districts\***



Rural districts generally gain population, while urban districts are more likely to lose population. House District 49 (Weld County) is 3.5 percent above the 2010 Census population, the rest of the districts are within 3 percent.

Source: Colorado Department of Local Affairs, State Demography Office. Available at: <http://arcg.is/1X4afz>  
 \*Difference in population totals between the 2010 Census and differential privacy demonstration file.

## Feedback on the April 2021 Census Demonstration Files

David Van Riper, Jonathan Schroeder, and Steven Ruggles  
IPUMS-NHGIS  
University of Minnesota  
May 28, 2021

### Introduction

Producing accurate, usable data while protecting respondent privacy are dual mandates of the U.S. Census Bureau. In 2018, the Census Bureau announced it would use a new disclosure avoidance technique based on differential privacy for the 2020 Decennial Census of Population and Housing. Instead of suppressing data and swapping sensitive records as the Bureau had previously done, the new approach injects noise into counts. Unfortunately, noise injection also makes the data less accurate and can hamper many use cases.

The Census Bureau has released five demonstration products that apply different versions of the new approach to 2010 census data. To assess the most recent demonstration products, we compare them with previous demonstration products and with the originally published 2010 summary data. The final two demonstration products were released on April 28, 2021. The e4 product has the same overall privacy budget ( $\epsilon \approx 4.4$ ) as previous demonstration products but reallocates the budget to different geographic units and modifies post-processing. The e12 budget is much larger ( $\epsilon \approx 12.2$ ), which would be expected to provide substantially greater accuracy. We understand that the e12 product uses the same parameters that the Census Bureau currently plans to use for the 2020 census.

### Analysis of Total Population Counts

We first analyze large discrepancies between the five demonstration products and the originally published 2010 data, measured as the percentage of geographic units where population counts differ by more than 5%. We compare four different geographic classifications: tracts, block groups, places, and American Indian Areas/Alaska Native Areas/Hawaiian Home Lands (here abbreviated as American Indian). We divide the population size of each geographic classification into deciles based on the 2010 total population.

Figure 1 shows the percentage of each decile of each geographic unit that differed from the true 2010 population counts by over 5%. The five rows of charts represent the five demonstration products produced by the Census Bureau. The leftmost column of charts show the error for census tracts. The demonstration products perform well for all but the smallest tracts, but the most recent demonstration products are not as accurate as the earlier ones.

The second column of Figure 1 shows the percentage of census block groups with error in total population counts exceeding 5%. These errors are substantially greater than the tract errors, and are especially pronounced for the April 2021 e4 product, which performs substantially worse than any prior demonstration product. By contrast, the accuracy of census places, shown in the third column of graphs, has improved substantially in the most recent demonstration products.

Finally, the population counts for American Indian Areas/Alaska Native Areas/Hawaiian Home Lands, although improved relative to the earliest demonstration products, remain problematic, with substantial error for all but the largest units.

Even where these charts show *relatively* few errors, such as for places in the April 2021 e12 product, there remain many instances of unacceptably large error. For example, the census-designated place of Fire Island, NY, had a 2010 population of 292, but the e12 product reports it as 392, a +34% error. The village of Vandalia, MI, has had a population between 300 and 450 in every census since 1880, including 2010 when its population was 301, but the e12 product reports its population as 245, a -19% error.

### **Analysis of Black and Hispanic/Latino Population Counts**

To understand how the disclosure avoidance measures affect the counts for population subgroups, we carried out the same analyses for the Black-alone population (Figure 2) and the Hispanic/Latino population (Figure 3). These figures show far more large discrepancies than the total population counts.

For the Black population shown in Figure 2, the new demonstration products do not represent a substantial improvement over prior releases, and the pervasive discrepancies are disturbing. For most block groups and places the discrepancy in the Black population exceeds 5% in every demonstration product, and even the e12 product--which ought to be the most reliable one--does not perform much better than earlier demonstration products. For the Black population the census tracts in the April 2021 e4 product perform no better than the previous releases, and even the e12 product is only a small improvement

The Hispanic/Latino population, shown in Figure 3, is even less accurate than the Black population. The discrepancies exceed 5% for the great majority of geographic units. Again, there is little or no improvement between the most recent data releases and the earlier ones, and even the E12 product indicates unacceptable levels of error.

### **Additional Errors and Inconsistencies**

Errors in other characteristics are equally problematic. The errors in counts of the number of children in the population of administrative units would wreak havoc on educational planning. For example, the e12 data product has a +3.4% overcount of children in Hoboken City School District, NJ (population 50,005); a +7.2% overcount in Naches Valley School District, WA (pop 8,078); and a +12.0% overcount in Mendocino Unified School District, CA (pop 5,665). Not only are there errors in the number of children in school districts, it appears that those errors include systematic biases. Among the smallest decile of school districts, 63% have an overcount of children, with a mean percent error of +6.1%. This is an example of a pervasive systematic bias found throughout these datasets: where counts are very small, they tend to be biased upwards.

In addition to the many large relative errors (mostly, though not all, in smaller counts), there are also numerous cases of very large absolute errors. In the e12 product, the total population of the Los Angeles School District is 5,950 above the actual count. Because of the large

population of the district, this overcount is only a 0.1% increase in its total population, but an overcount of nearly 6,000 is still not a “small” inaccuracy, and importantly, it is not a *uniform* overcount among all groups. Most of the increase is due to an overcount in *children* by 4,790, a more substantial 0.4% increase, and in a group of particular importance for a *school district*. Conversely, the E12 count of children in the neighboring Long Beach School District is low by 1,536, resulting in a -1.2% error in a very large school district (total population 510,940). Such inaccuracy is sure to have adverse impacts on these districts’ ability to serve their students and their families effectively.

The data also include numerous logical inconsistencies. For example, there are many cases in which the number of households exceeds the population size. Among county subdivisions, the largest such discrepancy occurs in Republican City township, NE (a minor civil division). The township actually had 131 residents and 67 households, but according to the E12 data product, it had 140 residents and 180 households. This is impossible, of course, but it also represents a +269% error in the household count as well as a +24 percentage-point error in the township’s housing occupancy rate (from 14% to 38%). Lewis town, VT, had no residents in 2010, but the E12 product assigns it 8 households (with no population). There are also many cases in which the number of adults in the population is implausibly low, including *91,047 blocks* where the E12 product codes *all* residents as children, with no adults present.

## **Conclusion**

The new demonstration products are limited to the content of PL94-171, and therefore do not permit analysis beyond a small number of very simple characteristics. Earlier demonstration products revealed major problems in other characteristics--such as age distributions--and we are unable to assess whether these errors have been addressed.

Given the limited time available and the limited content provided in the new demonstration products, we were unable to conduct more than a basic analysis. Nevertheless, that basic analysis yields profoundly disturbing results.

There were minimal improvements in the performance of the new demonstration files relative to the previous ones. We were disappointed to discover that the E12 file is not substantially more accurate on most measures than the e4 files. We were also dismayed to learn that the new datasets were virtually as bad as the previous ones with respect to the accuracy of counts for minority populations. The Census Bureau describes the e12 product as highly accurate. We find that although the e12 product has mitigated some egregious errors for the total population, major discrepancies remain for minority populations. This level of error will severely compromise demographic and policy analyses.

The demonstration files include troubling cases with extreme error in total or adult population counts, even if these are comparatively rare. Small localities can sometimes have their population doubled or halved by the disclosure avoidance noise. For example, the e12 product doubles the population of Saltaire village, NY, from 37 to 75, and it triples the population of Islandia city, FL, from 18 to 57.

For those who might object that these examples constitute cherry picking, we note that each “cherry” represents a community that deserves good data. The planned system would enter every community into a bad data lottery where the losers suffer for 10 years with material losses of federal funding. Litigation by undercounted communities is inevitable, and in these cases the Census Bureau will probably be forced to release the true counts.

Based on our analysis of the new demonstration products, we conclude that they are not ready for public release. We found pervasive biases and inconsistencies, high levels of inaccuracy in the counts of minority populations, and isolated large errors in the population counts for particular communities. Accordingly, the disclosure avoidance measures used in the e12 data product make the data unfit for many research and administrative purposes.

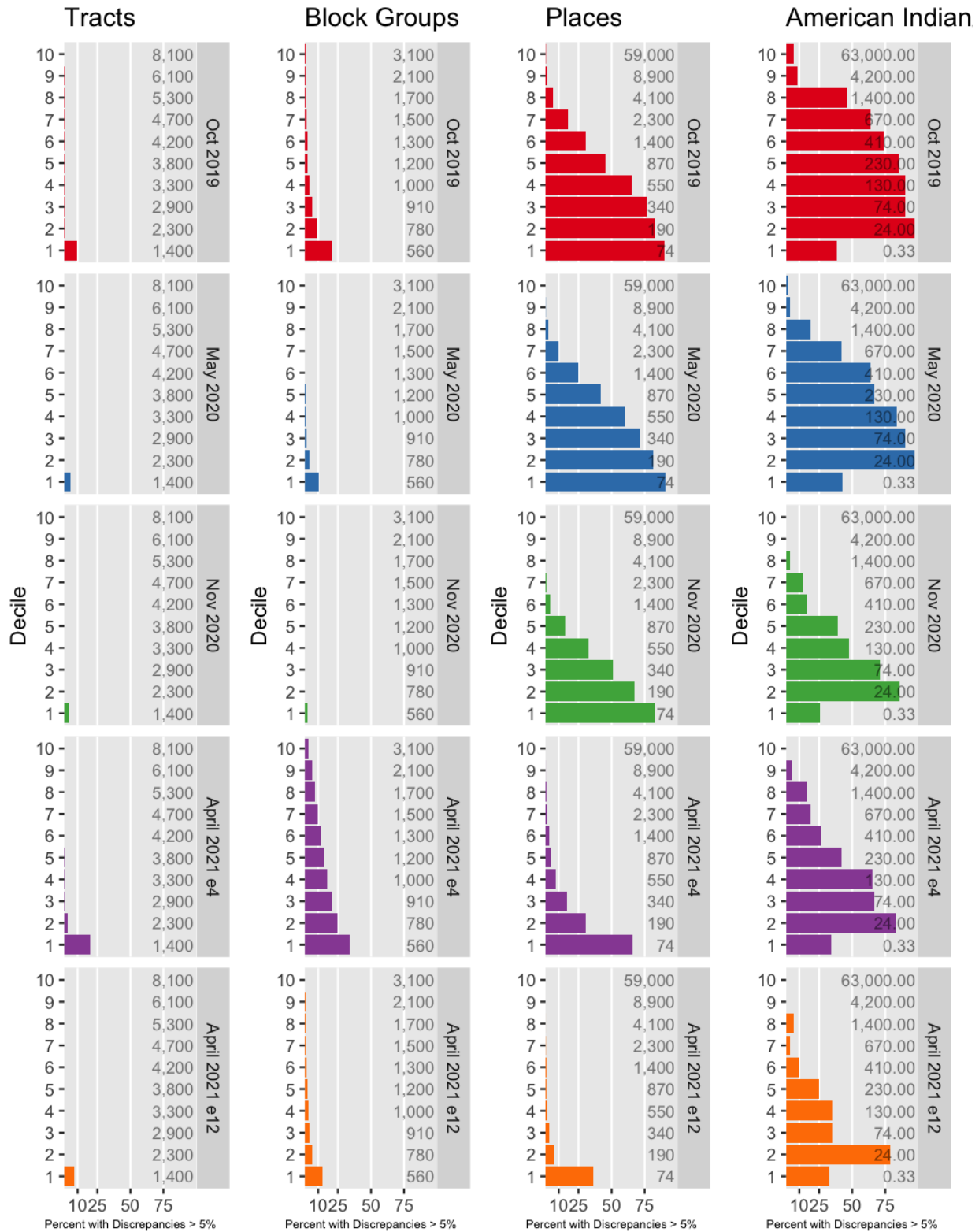


Figure 1. Percentage of units with a discrepancy between the demonstration data and 2010 Summary File 1 products greater than 5% for total population counts.

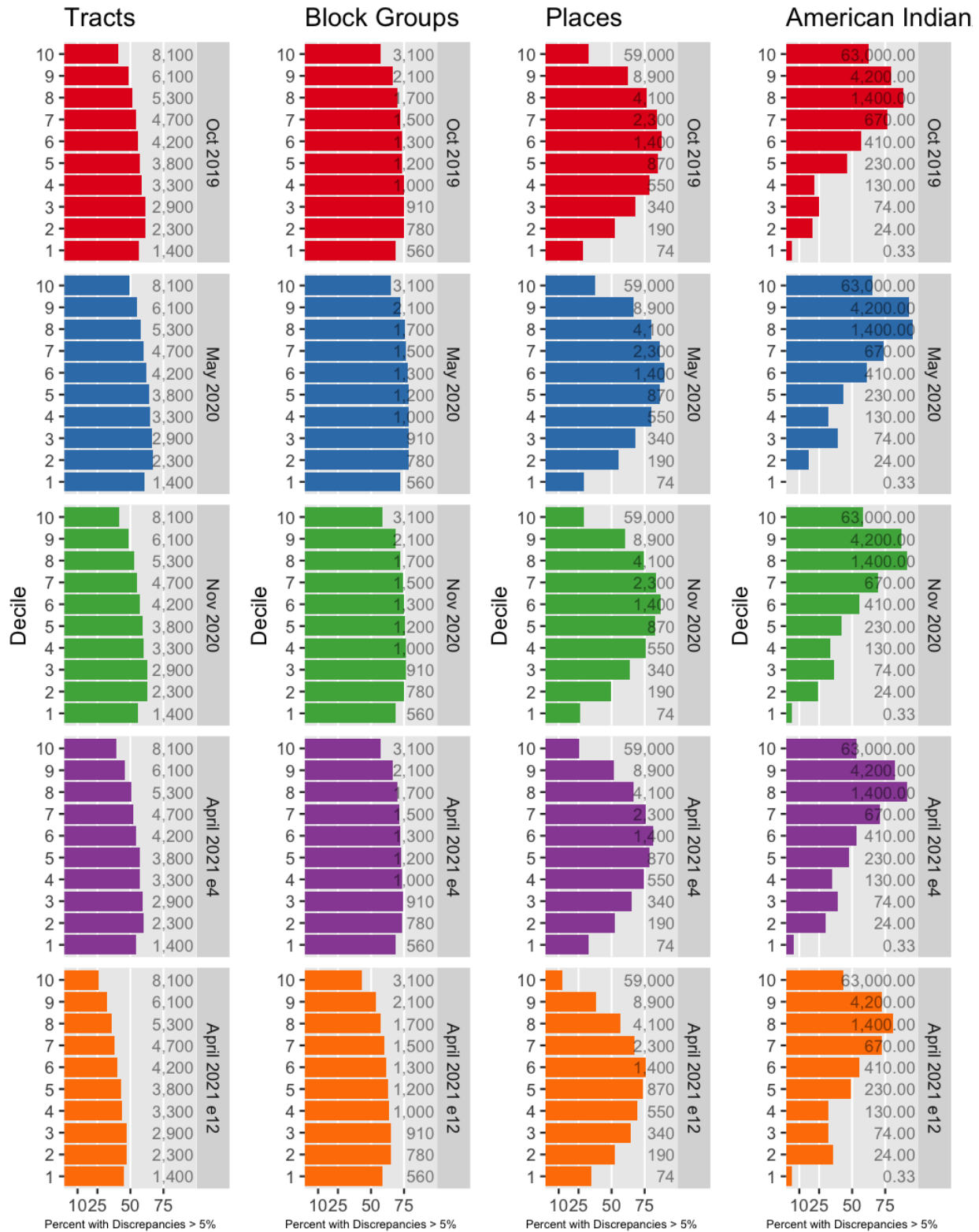


Figure 2. Percentage of units with a discrepancy between the demonstration data and 2010 Summary File 1 greater than 5% for Black-alone population counts.

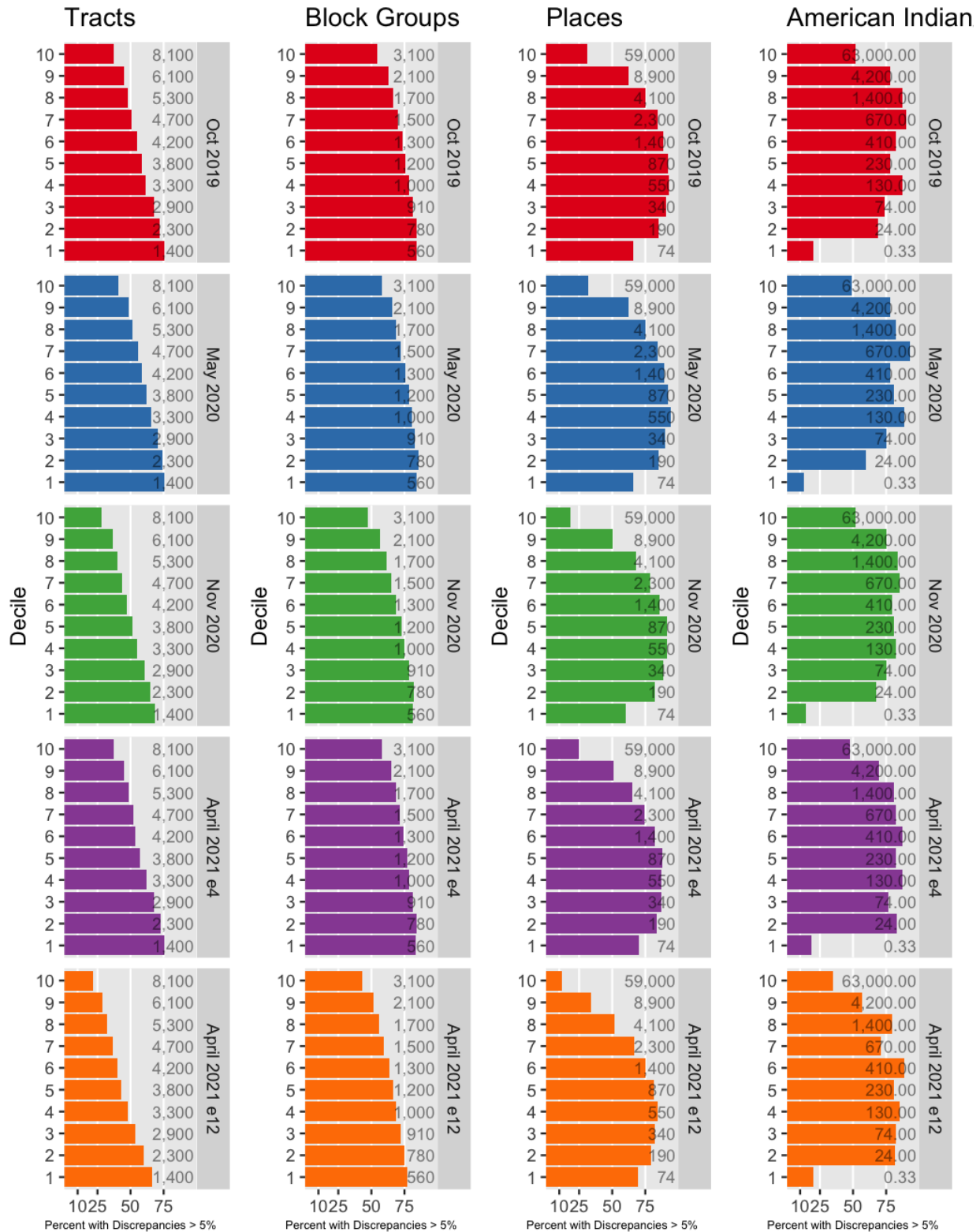


Figure 3. Percentage of units with a discrepancy between the demonstration data and 2010 Summary File 1 greater than 5% for Hispanic population counts.





STATE OF MAINE  
DEPARTMENT OF ADMINISTRATIVE & FINANCIAL SERVICES  
BURTON M. CROSS BUILDING, 3<sup>RD</sup> FLOOR  
78 STATE HOUSE STATION  
AUGUSTA, MAINE 04333-0078

SERVING THE PUBLIC AND DELIVERING ESSENTIAL SERVICES TO STATE GOVERNMENT

JANET T. MILLS  
GOVERNOR

KIRSTEN LC FIGUEROA  
COMMISSIONER

Dr. Steven Dillingham  
Director of the United States Census Bureau  
4600 Silver Hill Road  
Washington, DC 20233

February 20, 2020

Dear Dr. Dillingham,

The Office of the State Economist, within the Department of Administrative and Financial Services, serves as the State Data Center lead for the State of Maine. In this capacity, we are writing to express our concerns regarding the proposed policy changes involving the use of differential privacy in census data. Privacy protections for individuals are of utmost importance to the State of Maine. We recognize that caution and careful planning for disclosure avoidance are necessary in order to maintain the integrity of the decennial census and all Census products. However, upon careful review of the 2010 demonstration data product released by the U.S. Census Bureau, we are hereby voicing concern for the usability, reliability, and equity of differentially private (DP) Census data.

Our analyses show that small, rural places suffer the most in terms of inaccurate estimates. In Maine's case, that means a majority of our counties and sub-county geographies are subject to unacceptably high levels of error. If this holds true in the release of the 2020 decennial census data and other future data products, the repercussions for our state and nation are considerable.

Decennial census data are used for the apportionment of state legislative districts. They serve as the benchmark for population estimates, demographic projections, surveys, research, and analysis carried out by everyone from local housing planners to the U.S. Census Bureau itself. Over three hundred federal spending programs distribute funds on the basis of data derived from the decennial census. Policy decisions at all levels of government use data that originate with the decennial census. In many cases policymakers, researchers, businesspeople, and the public rely on data that is only available from the U.S. Census Bureau. If the reliability of that data falls by the wayside or the data becomes so difficult to interpret that general users are unable to decipher it, we run the risk of basing decisions on no data at all or, perhaps worse, on inaccurate data.

The U.S. Census Bureau has long been the standard-bearer in terms of providing high quality, reliable data to the public. This proposed policy change would threaten that position and throw into doubt any redistricting, funding decisions, or analysis done using census data.

**IRC\_00358**

While we have been able to assess the errors in the demonstration product, this will not be possible for the 2020 published data. At the time of writing, there is no established guidance with respect to how statistical analysis should be carried out in light of the proposed change. Even if these tools existed, we fear many of the data users within our state do not have the resources and training necessary to account for these errors. This exacerbates our concern that DP has the potential to exclude rural and resource-strained communities from equitable access to high-quality, reliable data, and that our narratives will be systematically misinformed as a result.

In light of our grave misgivings concerning this proposed policy change, we have several requests that would help to either reduce the negative impacts from the change or provide additional information to help us prepare for the impacts.

**1. We request that the U.S. Census Bureau release more demonstration datasets for different epsilon values, geographical hierarchies, and queries, as well as multiple iterations of each.**

**2. We request that the U.S. Census Bureau use a higher value of epsilon, and particularly higher allocation for Age and Sex tabulations.**

**3. We request that the U.S. Census Bureau release raw noise-injected counts.**

**4. We request better information and analysis from the U.S. Census Bureau regarding the impacts on related data products including the American Community Survey, Current Population Survey, and Population Estimates Program.**

**5. We request that the U.S. Census Bureau report margins of error or confidence intervals for previously released DP data and any newly-released DP data.**

Despite the availability of the demonstration data product, data users have not been given enough time to conduct thorough analysis to understand these impacts, since several tables were either not included or are not comparable to the demonstration data. For example, the U.S. Census Bureau has cautioned that table P20 is not comparable to the demonstration product. There has been inadequate opportunity to evaluate the privacy-accuracy tradeoff since there has been only one single demonstration data set to analyze at one single epsilon value, geographical hierarchy, and query. More demonstration datasets would allow users to understand these three important aspects of the privacy algorithm.

Additionally, there has been inadequate communication regarding impacts to other valuable data products such as the American Community Survey, the Current Population Survey, or the Population Estimates Program<sup>1</sup>. Other economic data released by the U.S. Bureau of Economic Analysis, U.S. Bureau of Labor Statistics, and a vast spectrum of other data agencies will similarly face challenges with survey design.

---

<sup>1</sup> The Census Bureau's analysis of the Population Estimates Program shows Maine (statewide) has the second-highest Mean Absolute Percentage Error (MAPE) among all states in these estimates: 42.5% MAPE using the demonstration products as a benchmark compared to 12.8% with published Census data. These estimates are a primary data input for Maine's population projections. Still, the data for this calculation has not been released to the public, which has left us mostly unaware of these impacts.

Inaccuracy in the decennial census will flow through ten full years of data via these crucial products. The current implementation of DP creates a group of regions and people, predominantly rural and already marginalized, that are left behind; they will continue to be left behind for the remainder of the decade unless action is taken to improve the algorithm. Without resolution to the above uncertainties it will be impossible to measure the magnitude of these errors, resulting in further challenges for these places and communities.

Following is a description of the analysis performed by our office and the results that prompted our concerns. We appreciate your consideration of our requests and look forward to a prompt reply.

Sincerely,



Angela Hallowell  
Maine State Data Center lead



Amanda Rector  
Maine State Economist

## Impacts in Maine

---

The demonstration data product was accessed courtesy of IPUMS NHGIS, University of Minnesota, [www.nhgis.org](http://www.nhgis.org). We find that most counts are reliable at the state level, as are total population counts at the county level. However, detailed counts for nearly all sub-state geographies have been compromised by noise injection.

### County-level counts

One example of this lies in age and sex counts at the county level (**Figure 1**). The greatest Mean Absolute Percentage Error (MAPE) is found for 18-19 years, 20 years, 21 years, and 85 years and over cohorts for both male and female. Even when aggregated by sex, MAPE is over 10% in all abovementioned cohorts except 18 and 19 years (**Figure 2**). This data has a major part to play in the analysis carried out by numerous state agencies. For example, the ongoing opioid crisis throughout the state disproportionately affects young men in rural counties. Inaccuracies of this magnitude in population counts could lead to under- or over-calculations of overdose rates and would make it difficult to statistically detect changes across time and space. This makes the management of this public health crisis a nearly impossible task. Additionally, Maine has the oldest median age and the highest percent of the population age 65 and older of any state in the U.S. The high level of inaccuracy with the 85 and over cohorts will make planning for our rapidly aging population increasingly complex.

Similarly, **Figure 3** demonstrates the inaccuracy in counts for households by age of householder. Again, the youngest category (householder aged 15-24) and the oldest categories (75-84 and 85 years and over) have the highest errors. This translates to errors that halve or double these populations in some of Maine's smallest counties (**Table 1**).

Race of householder in occupied units is also significantly flawed (**Figure 4**). All racial categories except *White alone* have MAPE over 25%. In fact, only two have MAPE under 100% (Two or more races and American Indian and Alaska Native). In Franklin County, the count of households with a black or African American householder was more than 11 times its published count (**Table 2**). Any changes in Maine's diversity at a county level will be incredibly difficult to statistically detect and will undoubtedly lead to misinformed narratives about demographic comparisons over time and space. These examples are just some of the many large errors we found in the data at the county level in Maine.

### County Subdivision and School District Counts

Data users will find county subdivision counts almost entirely useless given the current privacy loss budget level and allocation. Total population counts are relatively acceptable for large county subdivisions. Error is large for the smallest subdivisions, but meaningfully falls below 10% absolute percent error at about 900 people. However, this leaves about 236 of 533 Maine county subdivisions vulnerable to large miscounts. This is demonstrated in **Figures 5 and 6**.

Age and sex counts are severely affected by noise injection. **Figures 7 and 8** show the MAPE by age and sex cohort and counties, respectively. No category (other than total) has a MAPE under 50%, and many have MAPE well over 100% for both sexes. Similarly, half of the counties have MAPE across category and geographies above 100%; the lowest is in York at 49.8%. These errors are altogether unacceptable and if left unchanged, we will caution users against relying on any of these data.

This will have myriad financial and economic repercussions for the “winners” and “losers” that municipalities will randomly become. One significant example is funding for school districts. **Figure 9** shows the losses and gains in the school-aged population. School districts stand to lose significant portions of funding as a result of a faulty headcount. For example, RSU 34 (serving Alton, Bradley and Old Town) lost 422 students from its school-aged children count. In 2011, there were 290 students attending its Leonard Middle School<sup>2</sup>. This loss is akin to artificially removing the students from more than an entire school from its school district. Conversely, some lucky school districts such as Deer Isle-Stonington Community School District would see a 35% increase in its school-aged population.

It is important to note that these results are based on random draws; outcomes for Maine could be entirely different in another iteration of the algorithm. For this reason, we close by urging the U.S. Census Bureau to provide more demonstration datasets and to release raw noise-inject data that include negative counts. This will help data users approximate margins of error for the 2020 published data and assess how these errors will manifest in the future. Without this ability, we will cease to use most of the published decennial data and be forced to seek alternative data sources.

---

<sup>2</sup> Source: Maine Education Data Warehouse

# Tables and Figures

Figure 1. Mean Absolute Percent Error for age and sex, all counties in Maine

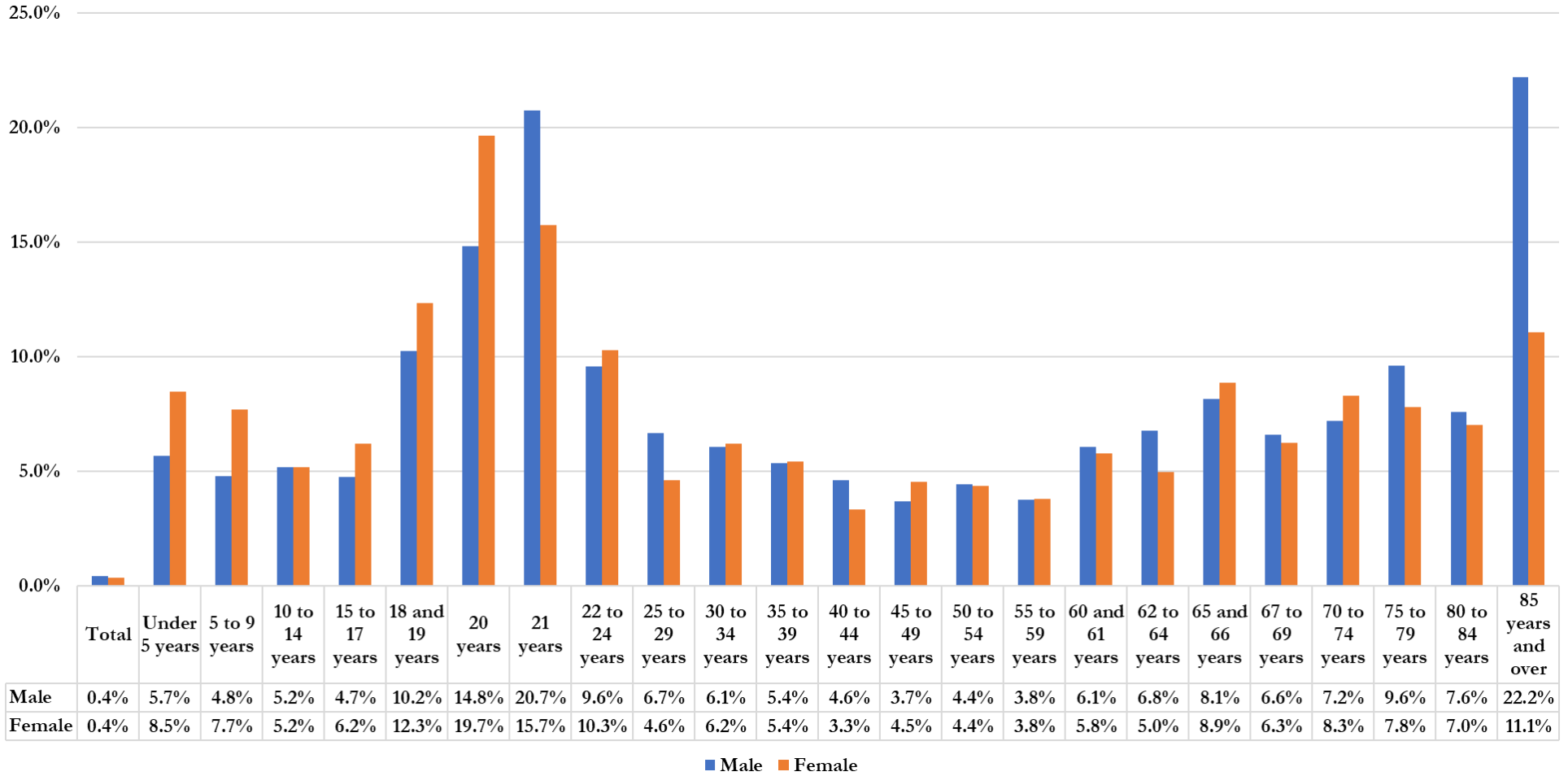


Figure 2. Mean Absolute Percent Error, both sexes, 16 Maine Counties

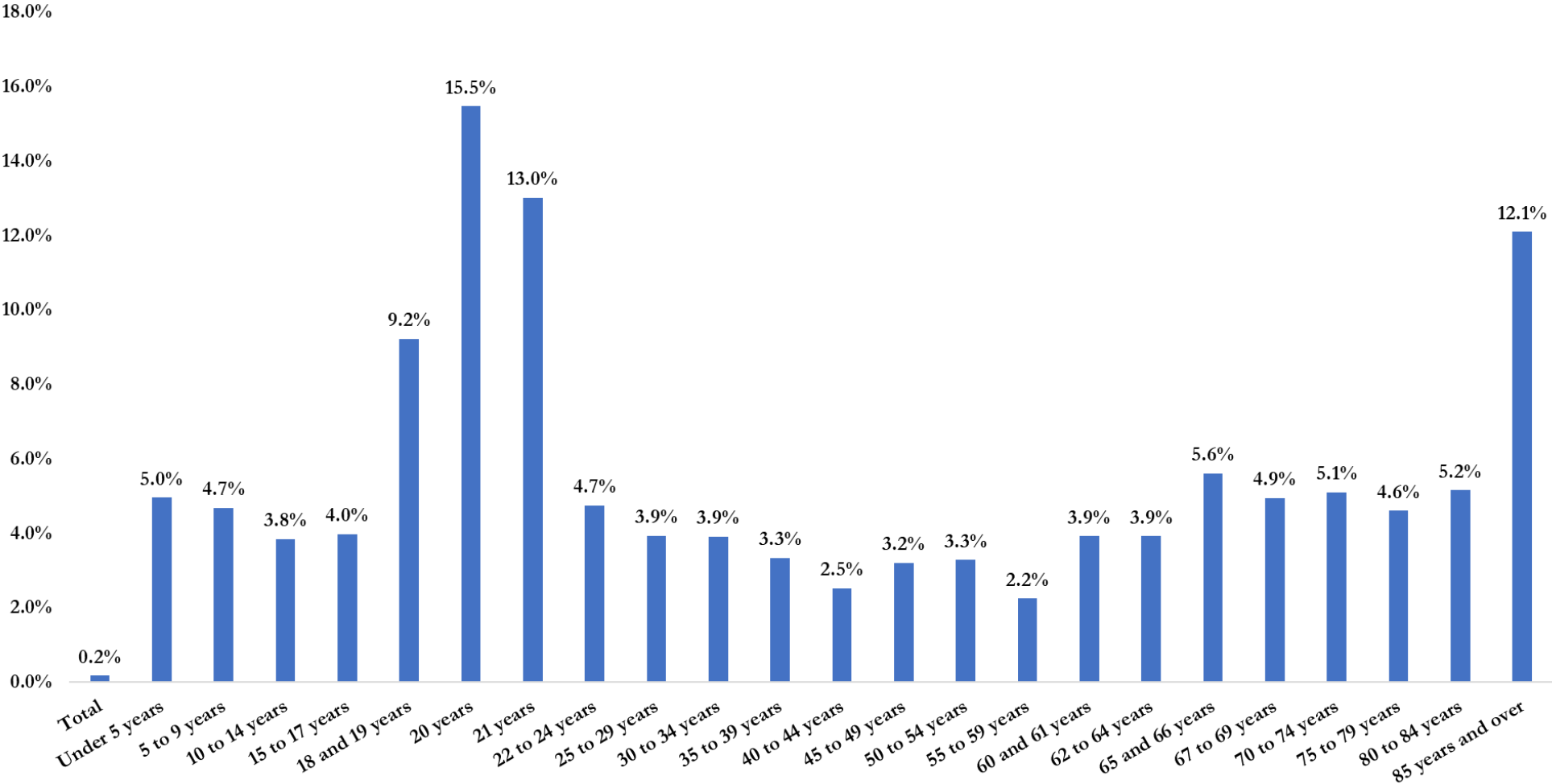


Figure 3. Mean Absolute Percent Error - Households by age of householder

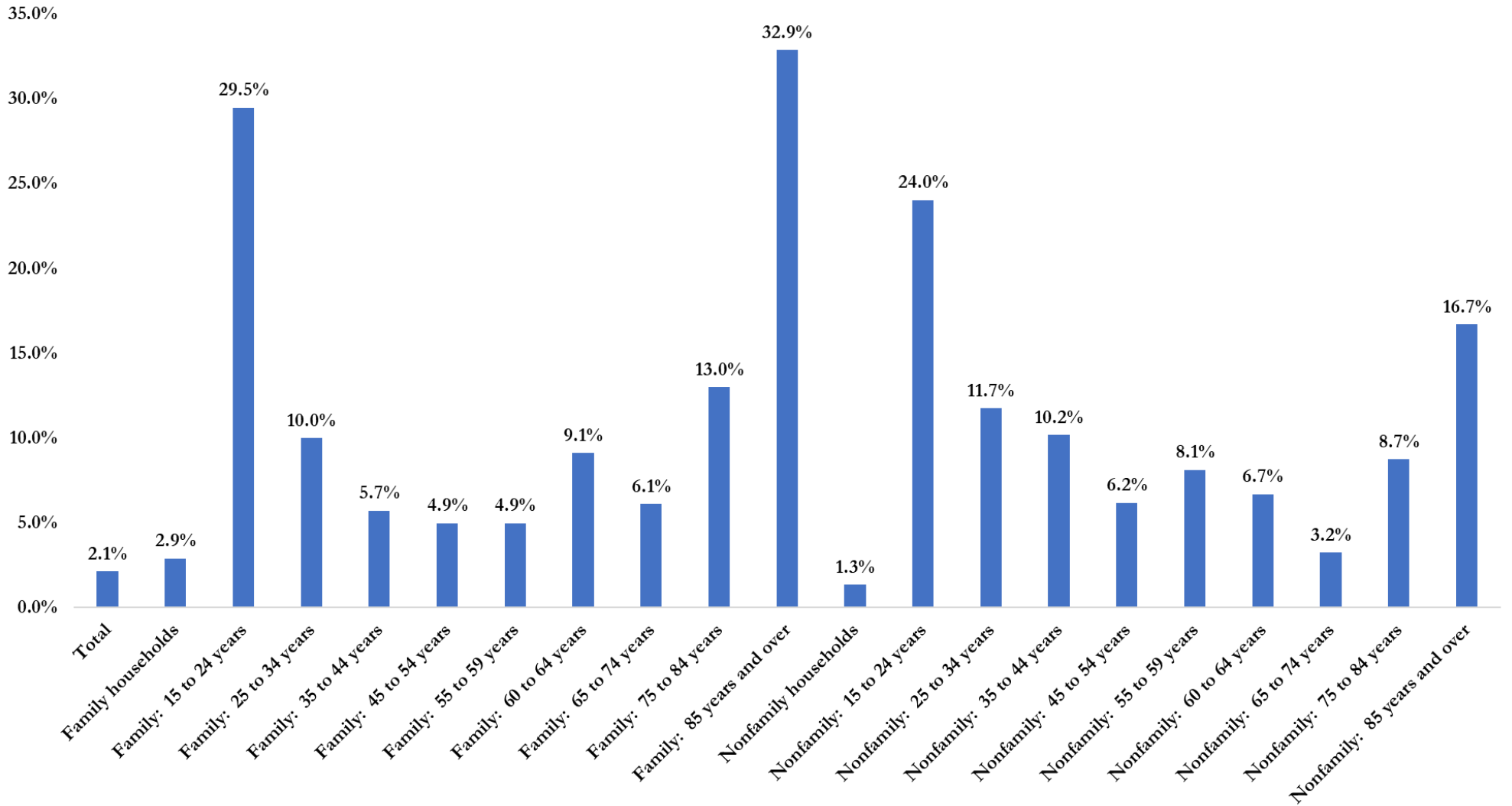




Figure 4. Race of Householder - Mean absolute percent error in 16 counties

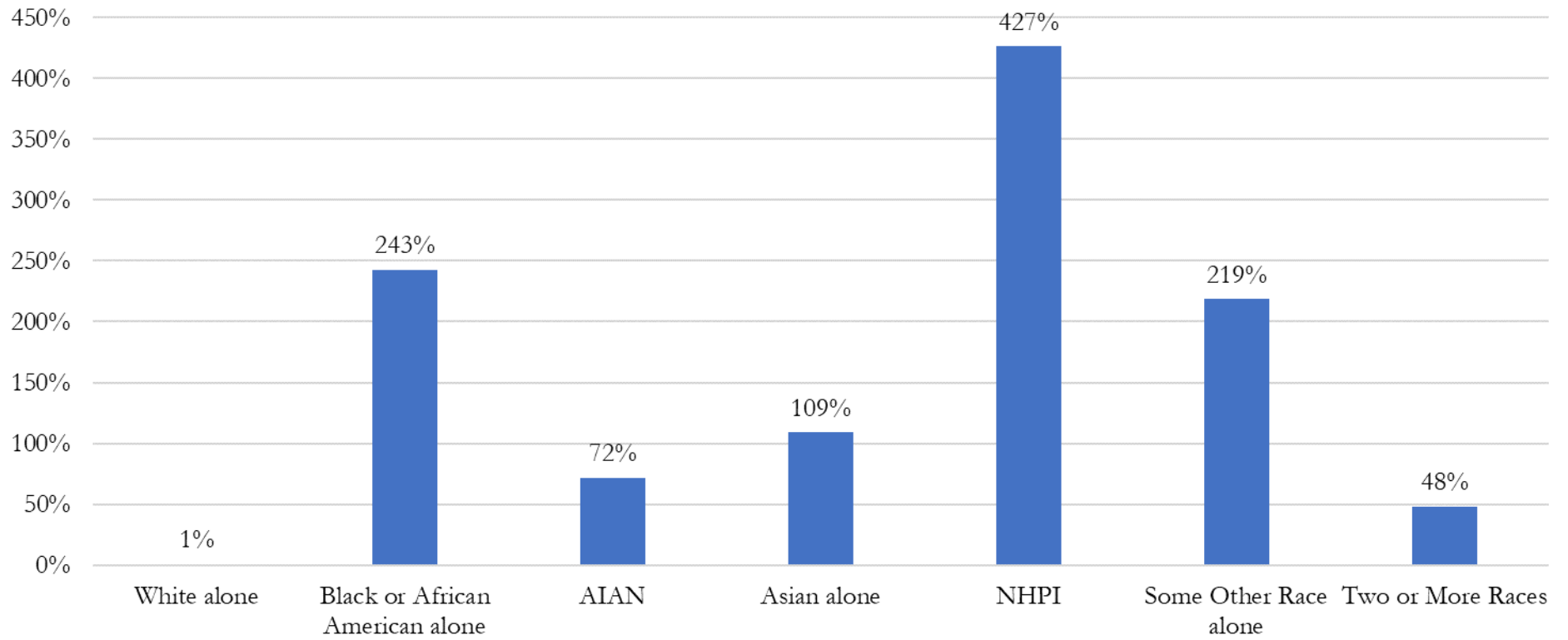


Figure 5. Percent Error in Total Population for All County Subdivisions

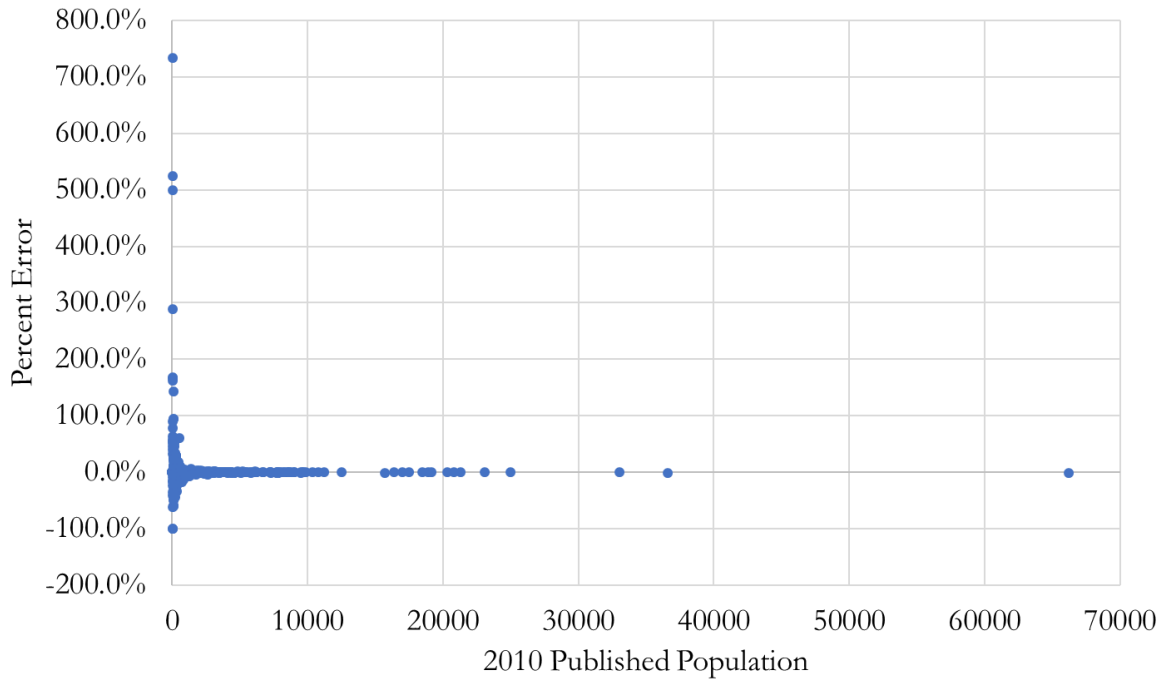


Figure 6. Percent Error in Total Population for County Subdivisions, Zoom View

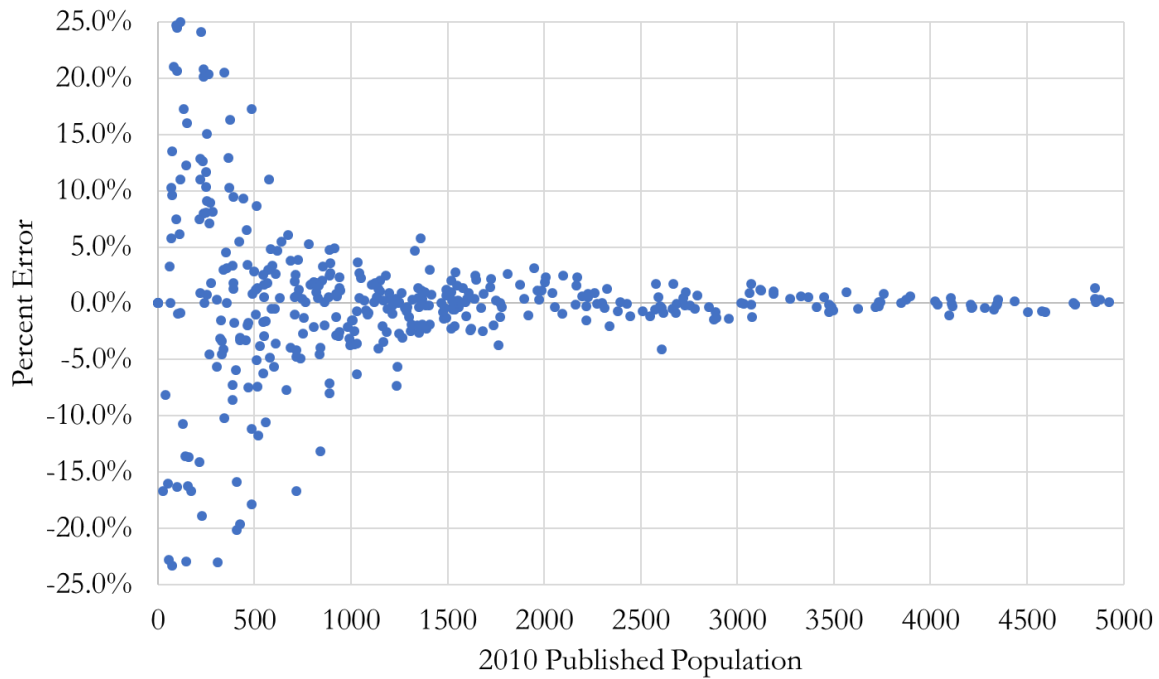
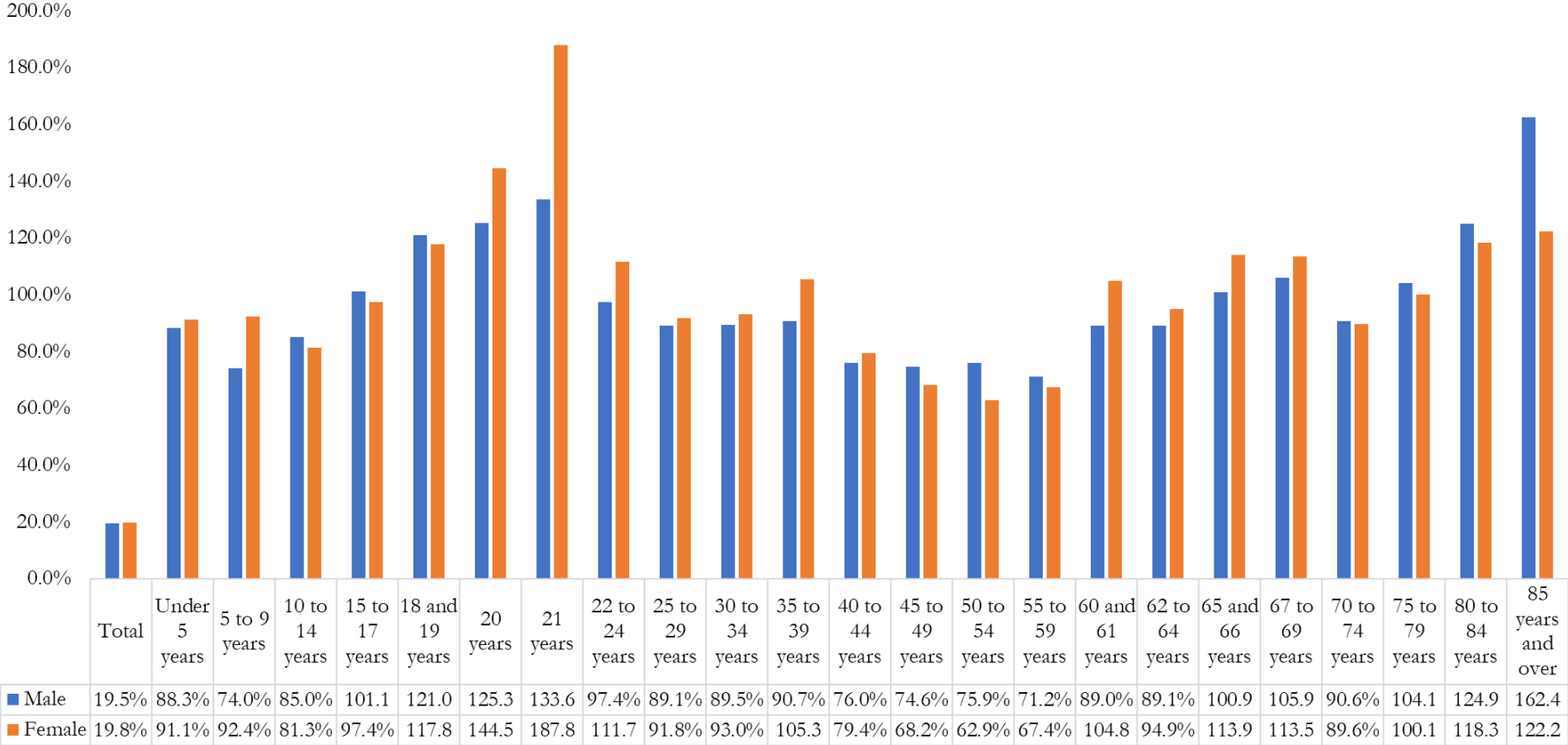


Figure 7. Mean Absolute Percentage Error for Age and Sex in Maine's County Subdivisions



■ Male ■ Female

Figure 8. MAPE for age and sex by county, all county subdivisions in Maine

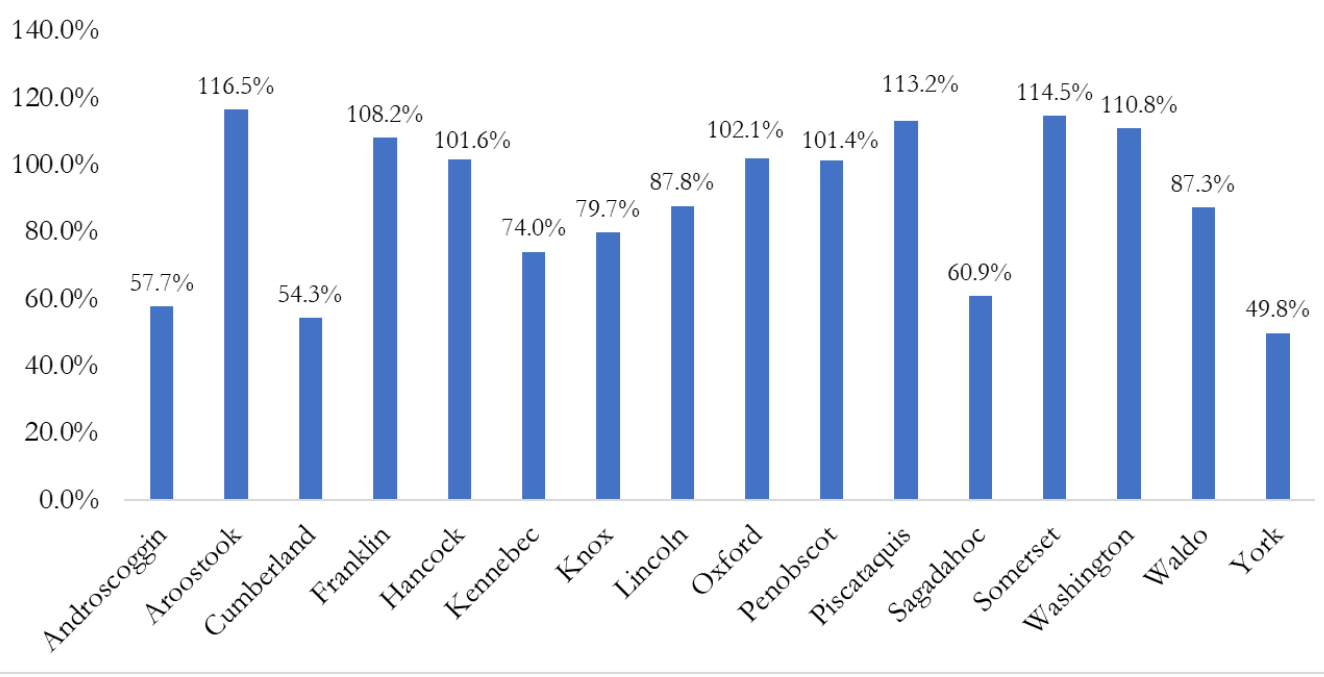
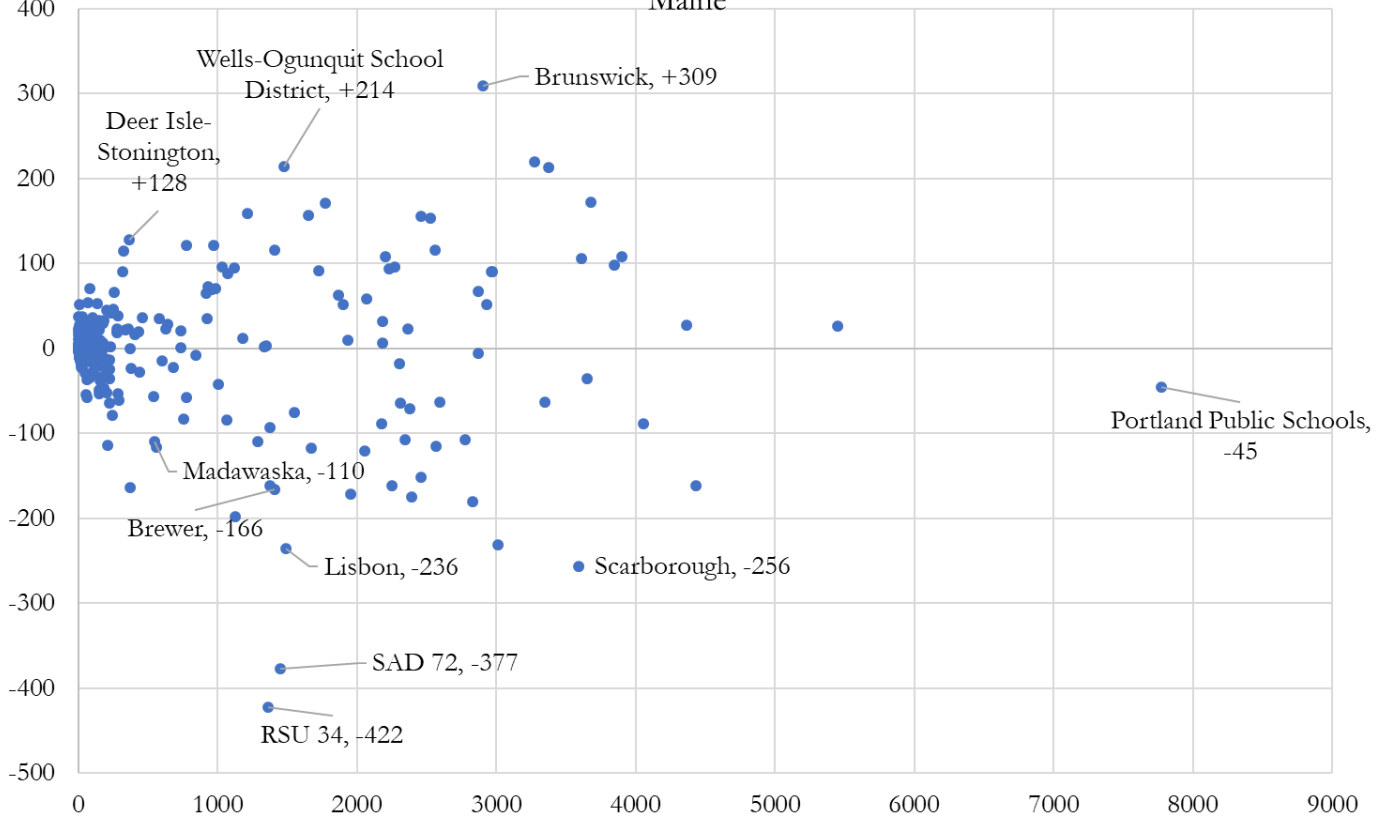


Figure 9. Miscount of school-aged children (5-17 years old) in all school districts in Maine



**Table 1. Households by type and age of householder – highest error categories by county**

**(Where 100% means doubling and -50% means halving)**

-50% and less | 
 -49% to -25% | 
 25% to 99% | 
 100% and over

Percent Difference	Family:	Family:	Family:	Family:	Family:	Family:
	Householder 15 to 24 years	Householder 25 to 34 years	Householder 45 to 54 years	Householder 60 to 64 years	Householder 75 to 84 years	Householder 85 years and over
Androscoggin County	-5%	3%	2%	7%	-2%	17%
Aroostook County	16%	-15%	8%	-1%	-6%	-23%
Cumberland County	20%	-2%	-1%	-3%	9%	-10%
Franklin County	70%	25%	2%	-11%	6%	-28%
Hancock County	11%	21%	7%	5%	-4%	28%
Kennebec County	-16%	-2%	0%	3%	-6%	-19%
Knox County	76%	17%	1%	8%	-12%	-37%
Lincoln County	109%	-20%	-3%	-1%	-13%	116%
Oxford County	2%	5%	0%	28%	-3%	-23%
Penobscot County	-20%	-5%	-1%	-2%	0%	10%
Piscataquis County	10%	0%	31%	29%	-51%	63%
Sagadahoc County	-2%	29%	1%	-6%	23%	-16%
Somerset County	16%	3%	-2%	-13%	14%	-4%
Waldo County	32%	-1%	-2%	-19%	47%	99%
Washington County	36%	-12%	17%	-10%	-11%	-31%
York County	-32%	1%	-1%	0%	-1%	3%

	Nonfamily:	Nonfamily:	Nonfamily:	Nonfamily:	Nonfamily:	Nonfamily:	Nonfamily:
	Householder 15 to 24 years	Householder 25 to 34 years	Householder 35 to 44 years	Householder 45 to 54 years	Householder 55 to 59 years	Householder 75 to 84 years	Householder 85 years and over
Androscoggin County	0%	-7%	10%	-6%	-1%	-6%	8%
Aroostook County	-7%	23%	2%	3%	-4%	-14%	5%
Cumberland County	-5%	-2%	1%	0%	-1%	-3%	1%
Franklin County	-8%	-2%	6%	-1%	-7%	10%	35%
Hancock County	-28%	-1%	1%	1%	7%	12%	-20%
Kennebec County	-6%	-9%	-4%	-6%	4%	5%	-2%
Knox County	-12%	-9%	-31%	6%	-1%	2%	27%
Lincoln County	33%	-11%	-21%	0%	6%	-1%	-4%
Oxford County	48%	21%	-20%	-9%	11%	-1%	15%
Penobscot County	-3%	-7%	11%	5%	0%	-1%	-11%
Piscataquis County	152%	14%	12%	28%	-45%	44%	-31%
Sagadahoc County	48%	39%	-4%	-17%	14%	5%	-39%
Somerset County	17%	-17%	-7%	2%	12%	9%	16%
Waldo County	-2%	17%	2%	-1%	-11%	-12%	7%
Washington County	4%	-5%	28%	12%	-2%	12%	-34%
York County	11%	-4%	-3%	-1%	-3%	-3%	12%

**Table 2. Percent Error for Race of Householder by County**

	<b>Total</b>	<b>White alone</b>	<b>Black or African American alone</b>	<b>AIAN</b>	<b>Asian alone</b>	<b>NHPI</b>	<b>Some Other Race alone</b>	<b>Two or More Races</b>	<b>MAPE</b>
Androscoggin County	0%	0%	-21%	23%	11%	-25%	72%	27%	<b>22%</b>
Aroostook County	1%	0%	82%	-30%	115%	-25%	627%	-11%	<b>111%</b>
Cumberland County	-1%	0%	-19%	24%	-8%	-33%	-58%	-15%	<b>20%</b>
Franklin County	5%	1%	1029%	182%	151%	260%	557%	81%	<b>283%</b>
Hancock County	1%	0%	105%	48%	-13%	1000%	26%	80%	<b>159%</b>
Kennebec County	-1%	0%	8%	-8%	-7%	92%	-53%	-20%	<b>24%</b>
Knox County	2%	0%	154%	137%	68%	500%	367%	16%	<b>155%</b>
Lincoln County	4%	2%	416%	23%	360%		364%	47%	<b>152%</b>
Oxford County	2%	1%	80%	118%	121%	25%	11%	22%	<b>47%</b>
Penobscot County	-1%	0%	0%	-16%	-15%	442%	9%	-36%	<b>65%</b>
Piscataquis County	8%	3%	813%	274%	423%	33%	683%	126%	<b>295%</b>
Sagadahoc County	4%	2%	73%	30%	23%	2400%	77%	162%	<b>346%</b>
Somerset County	2%	0%	186%	47%	140%	467%	462%	75%	<b>172%</b>
Waldo County	1%	0%	529%	118%	103%	480%	86%	-41%	<b>170%</b>
Washington County	1%	1%	367%	-29%	163%	1000%	0%	-8%	<b>196%</b>
York County	-1%	-1%	-7%	42%	-19%	-43%	-46%	-2%	<b>20%</b>
<b>MAPE</b>	<b>2%</b>	<b>1%</b>	<b>243%</b>	<b>72%</b>	<b>109%</b>	<b>427%</b>	<b>219%</b>	<b>48%</b>	



**NATIONAL REDISTRICTING  
FOUNDATION**

17 E. Monroe Street #214 • Chicago, IL 60603

**By Electronic Submission**

April 24, 2020

Honorable Steven Dillingham, Director  
U. S. Bureau of the Census  
4600 Silver Hill Road  
Washington, DC 20233

**Re: DAP2020**

Dear Director Dillingham:

I write today on behalf of the National Redistricting Foundation (“NRF”) to convey our significant concerns regarding the Census Bureau’s proposed use of differential privacy for the 2020 Census. We are concerned that the Bureau’s proposed application of differential privacy will substantially diminish the usability of the resulting data for redistricting, hampering the ability of state and local governments to comply with constitutional and statutory requirements that ensure fair and equal political representation. In particular, by generating inaccurate population and racial data for various geographies, the Bureau’s proposal risks undermining the constitutional principle of “one person, one vote” and the non-discrimination protections of the Voting Rights Act of 1965. Given the anticipated negative impact this new approach will have on drawing voting districts that accurately reflect and represent the people living in them, we urge you to reconsider—or at least recalibrate—your proposed approach.

The National Redistricting Foundation is a 501(c)(3) organization committed to preventing and reversing invidious gerrymandering, by promoting the public’s awareness of reapportionment and redistricting processes and engaging in legal action as appropriate to ensure that states’ redistricting and electoral processes result in fair representation. Bringing national attention to the importance of a fair redistricting process in 2021 is central to our mission, and elevating the need for a fair and accurate census in 2020 is a foundational piece of this work.

While we understand the need to protect data under Title 13 (U.S. Code) and to protect individuals’ information from being inadvertently disclosed, data accuracy is of paramount importance for redistricting at all levels of government. This fact must not be understated or under-appreciated. As things currently stand, we are deeply troubled by the sense that the privacy-loss budget tradeoff is unacceptably weighted against accuracy.

In particular, initial analyses based on the Bureau’s 2010 Demonstration Data Products suggest that the Bureau’s proposal risks undermining voters’ rights to equal and fair political representation in two ways.

National Redistricting Foundation is a 501(c)(3) tax-exempt organization. Contributions to National Redistricting Foundation are tax-deductible to the extent permitted by law.

**IRC\_00372**



## NATIONAL REDISTRICTING FOUNDATION

17 E. Monroe Street #214 • Chicago, IL 60603

**First**, the Bureau’s differential privacy proposal appears to generate significant inaccuracies in the population count for various geographic units within a given state. However, state and local governments rely on PL 94-171 redistricting data to reflect the *actual* count of persons residing in various geographic units. To ensure equal political representation, the United States Constitution, as well as various state laws, require state and local governments to redistrict based on defined—and often exacting—equal population standards. Under this constitutional principle of “one person, one vote,” congressional districts must not differ in population by more than one person. While state legislative districts have slightly more flexibility than congressional districts with respect to population deviation, the impact of differential privacy becomes more severe for smaller geographic units, where relatively low differences in population count can generate significant deviations from equal population requirements—wreaking havoc on the constitutional guarantee of equal political representation. These concerns would be compounded if, as some analyses indicate, the Bureau’s proposed algorithm tends to systematically redistribute population from urban areas to rural areas.<sup>1</sup> The resulting reallocation of political representation from urban communities to rural communities would do significant damage to the principle of political equality on which our constitutional democracy is based.

**Second**, accurate census data, down to even the smallest geographic level, is also essential to protect against gerrymandering, particularly at the expense of protected minority groups. The Voting Rights Act of 1965 (as amended), Section 2, provides a powerful tool to protect the voting rights of minority communities that have been historically, systematically oppressed. But the enforcement of Section 2 in redistricting is dependent on having accurate racial and ethnic data. The courts have made clear that Section 2 requires states to draw effective minority districts where, among other things, the minority group is able to comprise at least 50% of the district’s voting age population. The application of differential privacy may skew the data in minority districts, perhaps threatening their sustainability at or near the 50% minority voting age population requirement. In particular, initial analyses suggest that the Bureau’s differential privacy proposal can produce inaccurate counts for minority communities by reallocating population from larger minority groups to smaller ones and by geographically dispersing concentrated minority populations – precisely the kinds of inaccuracies that would work against the viability of majority-minority districts.<sup>2</sup>

As you continue to evaluate your options for applying differential privacy to the 2020 census, please make the necessary changes in your planning to maximize the extent to which the resulting data reflects the actual population counts, including with respect to racial groups, that

---

<sup>1</sup> See Memorandum from M. Gunter to Governor R. Northam, January 23, 2020, *available at* [https://www.ncsl.org/Portals/1/Documents/Redistricting/VA\\_CensusDistortionProgram\\_VAGovernor\\_2020-01-23.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/VA_CensusDistortionProgram_VAGovernor_2020-01-23.pdf); Letter from M. Mohrman to S. Dillingham, February 6, 2020, *available at* [https://www.ncsl.org/Portals/1/Documents/Redistricting/WA\\_OFM\\_DAS\\_Response\\_Letter.pdf](https://www.ncsl.org/Portals/1/Documents/Redistricting/WA_OFM_DAS_Response_Letter.pdf).

<sup>2</sup> See *supra* note 1.

National Redistricting Foundation is a 501(c)(3) tax-exempt organization. Contributions to National Redistricting Foundation are tax-deductible to the extent permitted by law.





**NATIONAL REDISTRICTING  
FOUNDATION**

17 E. Monroe Street #214 • Chicago, IL 60603

are enumerated for all geographic units within a state. Anything less threatens to undermine the constitutional principle of equal representation—and the rights guaranteed by the federal Voting Rights Act.

Thank you for your consideration of this comment and request.

Respectfully submitted,

Marina Jenkins  
National Redistricting Foundation

National Redistricting Foundation is a 501(c)(3) tax-exempt organization. Contributions to National Redistricting Foundation are tax-deductible to the extent permitted by law.

**IRC\_00374**



NATIONAL CONGRESS OF AMERICAN INDIANS

## POLICY RESEARCH CENTER

May 2021



### Research Policy Update *2020 Census Disclosure Avoidance System: Potential Impacts on Tribal Nation Census Data*

The purpose of this research brief is to review the recent U.S. Census Bureau Disclosure Avoidance System Demonstration Products that illustrate how privacy methods may be implemented on the 2020 Census data to protect confidentiality and review analysis of the potential impacts on Tribal Nation census data.

If you are new to the U.S. Census Bureau privacy measures topic (Disclosure Avoidance System, Top Down Algorithm, Differential Privacy), we recommend first reviewing the following Research Policy Updates to learn the basics and gain a solid background before reading this update:

- [Differential Privacy and the 2020 U.S. Decennial Census: Impact on American Indian and Alaska Native Data](#) (2019)
- [Decennial Census: Key Uses of the Data](#) (2020)
- [Differential Privacy and the 2020 Census: A Guide to the Data and Impacts on American Indian/Alaska Native Tribal Data](#) (2021)

#### **Census Privacy Methods – Introduction to the Demonstration Products**

The U.S. Census Bureau says it is committed to protecting the private information collected through any of the U.S. Census Bureau surveys or censuses that identify an individual or business.<sup>1</sup> The U.S. Census Bureau says it is not only committed to protecting individual privacy, but is prohibited by law (Title 13) from disclosing or publishing “any private information that identifies an individual or business, including names, addresses (including GPS coordinates), Social Security Numbers, and telephone numbers.”<sup>2</sup>

The commitment to maintain the confidentiality of individual data and concerns about third party re-construction and re-identification of public census data on individuals led to the planned use of new privacy methods for the 2020 Decennial Census dataset and tabulations. The Census Bureau has produced several demonstration products, called Privacy-Protected Microdata Files (PPMFs), to allow data users to see the impacts of the new privacy methods and various changes in the algorithms they use to process the Census data.<sup>3</sup>

The demonstration products test adjustments to the new privacy measures planned for 2020 Census data by using these methods on 2010 Decennial Census data. This allows data users to compare the demonstration products or tests of the algorithm changes to the 2010 census data in the 2010 Summary File 1 dataset to see the potential impacts of the privacy methods on the accuracy and usability of census data. **Figure 1** shows the release dates of the six publically available test results/demonstration products.<sup>4</sup>

**Figure 1. Census Demonstration Products**

Demonstration Product/PPMF	Date Released
Demonstration 1	October 29, 2019
Demonstration 2	May 27, 2020
Demonstration 3	September 17, 2020
Demonstration 4	November 16, 2020
Demonstration 5 – PLB 12.2	April 28, 2021
Demonstration 6 – PLB 4.3	April 30, 2021

### April 2021 Demonstration Products – *What was Produced?*

The U.S. Census Bureau released two new demonstration products at the end of April 2021. Each demonstration product released to the public provides a glimpse into what might happen with the accuracy and usability of the 2020 Census data when the privacy system is applied in a certain way.

Differential Privacy allows the application of more accuracy in certain parts of the census dataset by allocating a “Privacy Loss Budget” to data for priority uses. The first five demonstration products kept a similar level of accuracy in the data i.e. the Privacy Loss Budget (PLB) was kept at the same level that had a high level of privacy protection and lower accuracy. By keeping the PLB at the same level for the first five demonstration products, data users could focus on how changes in the algorithm and geography hierarchy impacted the data accuracy rather than the changes from an increase or decrease of PLB.<sup>5</sup> The privacy methods include changes in the algorithm used, the census geographies, and the level of the PLB, which all impact the data quality, accuracy, and usability.

The two new demonstration products released in April 2020 show the impacts to the data from recent changes in the privacy algorithm and Privacy Loss Budget. Demonstration product six (Demonstration 6 – PLB 4.5) shows the recent impacts of changes in the algorithm and geographies from demonstration products one through four. Demonstration product five shows how an increase of the Privacy Loss Budget in addition to those changes impact the accuracy of the data (Demonstration 5 – PLB 12.2).<sup>6</sup> Both April 2020 demonstration products use the same algorithm and geographic hierarchy, but the U.S. Census Bureau has indicated that the PLB of 12.2 is similar to the level of privacy and accuracy they will produce in the final 2020 Census data.

## AI/AN Census Data – Analysis of Demonstration Products

The National Congress of American Indians (NCAI) Policy Research Center analyzed the April 2021 Census Demonstration Products to identify impacts from the changes made to the algorithm and geographies (Demonstration 6 – PLB 4.5) and to determine how the increase in Privacy Loss Budget (Demonstration 5 – PLB 12.2) impacts the American Indian/Alaska Native tribal geography data.<sup>7</sup>

The U.S. Census Bureau produced Privacy-Protected Microdata Files (PPMFs) for the demonstration products and the University of Minnesota IPUMS National Historical Geographic Information System (NHGIS) tabulated the data into tables for 2010 Summary File 1 comparisons.<sup>8</sup> The NCAI Policy Research Center used the American Indian Area/Alaska Native Area/Hawaiian Home Land (by State-County-Census Tract) datasets from both April 2021 Demonstration Products on the IPUMS website.<sup>9</sup> This data is free and available for anyone to use.

The dataset includes the ability to analyze the data by the following census geographies: Federal American Indian Reservations/Off-Reservation Trust Lands, Alaska Native Village Statistical Areas (ANVSA), Oklahoma Tribal Statistical Areas (OTSA), State Reservations, Tribal Designated Statistical Areas, and State Designated Tribal Statistical Areas. The dataset also includes Hawaiian Homelands but this data was removed from our analysis to keep the focus of the analysis on the AI/AN Tribal Geographies. The NCAI Policy Research Center looked at 617 Census AI/AN tribal geographies representing tribal lands for both datasets.

The NCAI Policy Research Center prepped the datasets by identifying the AI/AN geographies or tribal land population sizes by categories and calculated the AI/AN Alone and In-Combination data for these tribal lands. **Figure 2** shows how many of the 617 tribal lands examined were within different population size categories. This is significant because the accuracy targets created by the U.S. Census Bureau for the datasets “ensured that the largest racial or ethnic group in any geographic entity with a total population of at least 500 people is accurate to within five percentage points of their enumerated value at least 95 percent of the time.”<sup>10</sup> Most AI/AN Census

**Figure 2. Tribal Lands by Population Size**

Population Size	Total Number of AI/AN Tribal Lands
< 500	353
500 – 999	83
1,000 – 2,499	48
2,500 – 4,999	45
5,000 – 9,999	36
10,000 – 24,999	27
25,000 – 49,999	10
50,000 – 99,999	5
100,000+	10

tribal lands have a population less than 500, which means that most tribal lands were not considered during the accuracy targets in the latest privacy methods.

The analysis sought to clarify error metrics produced by the U.S. Census Bureau and to understand the impacts of the privacy methods on AI/AN tribal land data. The analysis focused on the geography and race data but impacts on other characteristic data such as age and sex are also likely.<sup>11</sup> Impacts on AI/AN population data not on AI/AN tribal lands has also been shown to experience a negative impact on accuracy of the data in previous demonstration products.<sup>12</sup> This analysis only looks at data from the tribal geographies, which are referred to throughout this update as tribal lands.

### **April 2021 Demonstration Products – Impacts on AI/AN Tribal Lands**

The U.S. Census Bureau produced error metrics to help data users evaluate the quality of the census data following the demonstration product releases. The metrics often focus on the mean and absolute values of changes in the counts for particular geographies or population groups in the privacy protected data.<sup>13</sup> While these can help determine the progress of the changes in how privacy methods are applied among demonstration products, the error metrics don't show the full picture since they don't include standard deviations or ranges. An absolute error of 5 percent could mean an increase or a decrease, and even if the average result is low, there could be large changes in the data that is used to calculate the average. For example, a mean absolute change of five individuals may seem low, but the range of the data could show significantly more individuals either gained or lost from different the different tribal lands.

The Census privacy protections are applied through statistical methods that create errors in the data to promote privacy. Each time the privacy protections are applied, the impact is random, and some AI/AN tribal geographies or lands may actually end up with counts lower than the actual count (negative counts), higher than the actual count (positive counts) or even zero counts even though they had a population in the raw census data. **Figure 3** shows the total number of AI/AN lands that lost 100 percent of their 2010 Census total population, 100 percent of the AI/AN Alone population, and 100 percent of the AI/AN Alone and In-Combination population in the two April 2021 Demonstration Products as a result of the privacy measures applied to the data.

**Figure 3. Complete Population Loss in AI/AN Areas that had a Population above Zero in the 2010 Census Data (Summary File 1) – April 2020 Demonstration Products 6 (PLB 4.5) and 5 (12.2)**

Population	Number of AI/AN Tribal Lands with 100% Population Loss (PLB 4.5)	Number of AI/AN Tribal Lands with 100% Population Loss (PLB 12.2)
Total Population	3	2
AI/AN Alone Population	10	4
AI/AN Alone and in Combination	7	2

**Figure 3** shows that in both demonstration products with the lower and higher Privacy Loss Budget, there were tribal lands that lost their entire population. All of the AI/AN tribal lands that lost population were in the population size category of less than 500 people, and all were less than 15 people. These extremely small AI/AN tribal lands went from having a population to having no population after the privacy methods were applied. Some AI/AN tribal lands lost their entire AI/AN Alone Population and their entire AI/AN Alone and In-Combination Population after the privacy methods were applied. Other AI/AN tribal lands that experienced a complete loss of population in only one category (AI/AN Alone or AI/AN Alone and In-Combination) still saw extreme losses in the other category. The increase in Privacy Loss Budget between the Demonstration products six (PLB 4.5) and five (PLB12.2) shows improvement resulting in fewer AI/AN tribal lands losing the entire population, but some remain in that category.

**Figure 4** shows the number of AI/AN tribal lands that lost significant levels of population as a result of the privacy protections in both demonstration products (PLB 4.5 and PLB 12.2). The rows in Figure 4 are the number of AI/AN tribal lands that lost a specific percent of their population in the demonstration products due to the application of privacy methods compared to the reported population in the 2010 Summary File 1.

The large columns show the type of population lost by the AI/AN tribal lands. The column titled “Total Population” shows how many AI/AN tribal lands lost a certain percent of their 2010 population, or by how much their population counts decreased with privacy protections. If an AI/AN tribal land had a 100 percent population loss, there is no longer a population that exists on that AI/AN tribal land for that demonstration product. For the column AI/AN Alone, any losses mean that the AI/AN tribal land lost population that racially identified in the 2010 Census as AI/AN Alone. This could mean that the individuals who responded with AI/AN Alone racially became AI/AN In-Combination or it could mean that those individuals were no longer AI/AN at all in the data. This is why it is also important to note the changes in the AI/AN Alone and In-Combination column.

**Figure 4. Number of Census AI/AN Tribal Lands with Percent Population Losses in the April 2021 Demonstration Products in Demonstration Product 5 (PLB 12.2) and 6 (PLB 4.5)**

Percent Population Loss	Total Population		AI/AN Alone Population		AI/AN Alone and In Combination Population	
	PLB 4.5	PLB 12.22	PLB 4.5	PLB 12.2	PLB 4.5	PLB 12.2
100%	3	2	11	6	6	2
50 – 99.99%	3	0	11	9	7	4
25 – 49.99%	5	6	23	12	16	10
10 – 24.99%	30	10	64	33	58	32
5 – 9.99%	38	23	42	44	40	36
2 – 4.99%	67	54	75	55	69	54
>0 – 1.99%	140	170	91	125	89	112
≥10%	41	18	109	60	87	48
≥ 5%	79	41	151	104	127	84

**Figure 4** illustrates that the increase in Privacy Loss Budget (PLB) between the demonstration products from 4.5 to 12.2 reduced the number of AI/AN tribal lands with a population loss for each percent category. Although the number of AI/AN tribal lands for each percentage of population loss decreased, **Figure 4** doesn't show how much the total number of individuals lost on AI/AN tribal lands increased or decreased. For example, while some AI/AN tribal lands lost less than two percent of their population for both data products, some may still have lost hundreds or thousands of individuals from their population.

**Figure 5** provides a first look into the range of total individuals lost or gained in an AI/AN tribal geography after application of the privacy methods to the demonstration product data. **Figure 5** shows the maximum and minimum population gains and losses in AI/AN tribal geographies for both demonstration products. The three main columns show the losses for both demonstration products for the total population lost, the loss of population that identified in the 2010 Census as AI/AN Alone, and the loss of population of anyone in the geography who identified as AI/AN either alone or in combination with another race category.

**Figure 5. Range of Total Counts Gained or Lost in Census Data and April Demonstration Products (PLB 4.5 and 12.2) – Minimum and Maximum Values for Population Count Changes with Privacy Protections**

	Total Population for AI/AN Lands		AI/AN Alone Population for AI/AN Tribal Lands		AI/AN Alone and in Combination Population for AI/AN Tribal Lands	
	PLB 4.5	PLB 12.2	PLB 4.5	PLB 12.2	PLB 4.5	PLB 12.2
<b>Largest Count Lost (Min)</b>	-1107	-242	-198	-99	-593	-142
<b>Largest Counts Gained (Max)</b>	1047	254	229	223	422	239

The maximum and minimum values in **Figure 5** show that the extreme losses and gains from the demonstration produce six (PLB 4.5) lessen when the privacy budget increases in demonstration product five (PLB 12.2). Demonstration product six (PLB 4.5) showed the largest total population loss by an AI/AN tribal land was 1,107 individuals, the largest AI/AN Alone population loss by an AI/AN tribal land was 198, and the largest AI/AN Alone and in Combination population by an AI/AN tribal land was almost 600 individuals. Although, as shown in Figure 4, these may be small percent losses for some AI/AN tribal lands, these losses of population and AI/AN individuals on AI/AN tribal lands are significant if these numbers are used for local tribal governance, federal funding formulas, research, redistricting, and other uses. The losses in demonstration product six (PLB 4.5) are less extreme but still significant and question the basic usability of data at that level of privacy for the tribal lands that lost counts. While there are also some tribal lands that gained counts, that also means that the data is inaccurate and can also impact the uses stated above. The U.S. Census Bureau’s Disclosure Avoidance System with the use of Differential Privacy seems to create winners and losers among AI/AN tribal lands – some gain counts, some lose counts – in a random manner that mostly disadvantages small, rural, and remote populations.

In November 2020, the U.S. Census Bureau Data Stewardship Executive Policy (DSEP) Committee made the decision to not set the sum of AI/AN tribal geography counts in a state as invariant, or equal to the actual counts, at the state level.<sup>14</sup> Setting the AI/AN tribal land populations invariant at the state level would have meant that if someone added together all the population gains and losses after application of the privacy methods in AI/AN tribal lands within the same state, the total count would equal the accurate number of people counted in the Decennial Census.<sup>15</sup> This would not mean that each AI/AN tribal land has a true count. The sum of the privacy protected counts in all AI/AN tribal geographies in a state would equal the actual total count. It is unclear why the DSEP decided to continue to hold the overall state population invariant, but not to hold the sum of the AI/AN tribal geography counts invariant. Requests by tribal leaders to make each AI/AN tribal geography count invariant so that all



Tribal Nations could have accurate census data have not been adopted by the U.S. Census Bureau.

**Figure 6** shows the consequences of removing the population invariant for AI/AN tribal lands at the state level in demonstration products 5 and 6 by illustrating the total gains and losses in counts in all AI/AN tribal lands. **Figure 6** compares the population gains and losses on AI/AN tribal lands in both demonstration product five (PLB 12.2) and demonstration product six (PLB 4.5). The columns show the comparison of the two demonstration product gains and losses for the AI/AN tribal land total population, AI/AN Alone population, and AI/AN Alone and In-Combination population.

**Figure 6. Total AI/AN Losses and Gains in Counts – April Demonstration Product 5 (PLB 4.5) and 5 (12.2)**

Losses and Gains for AI/AN Tribal Lands	Total Population		AI/AN Alone Population		AI/AN Alone and in Combination Population	
	PLB 4.5	PLB 12.2	PLB 4.5	PLB 12.2	PLB 4.5	PLB 12.2
<b>Total Number of Tribal Lands (%) with Lost Counts</b>	286 (46.4%)	265 (42.9%)	317 (51.4%)	282 (45.7%)	289 (46.8%)	250 (40.5%)
<b>Total Number of Lost Counts in all Tribal Lands</b>	-10,455	-4,780	-5,798	-2,617	-6,554	-3,064
<b>Total Number of Tribal Lands (%) with Gained Counts</b>	286 (46.4%)	280 (45.4%)	253 (41%)	258 (46.7%)	288 (46.7%)	313 (50.7%)
<b>Total Number of Gained Counts in all Tribal Lands</b>	+8011	+3,685	+4,273	+2,236	+2,858	+4,204
<b>Overall Gain/Loss in Counts in Tribal Lands</b>	-2,444	-1,095	-1,525	-381	-291	+1,140

Percentages of Total Loss and Gain do not equal to 100 percent because AI/AN tribal geographies with a zero percent change, including those with zero counts in 2010, were not included in either the gain or loss calculation.

The two blue rows show the percent of the 617 AI/AN tribal lands examined with either a population gain or a population loss. The findings show that of the 617 AI/AN tribal lands examined, the percent of AI/AN tribal lands with population gains and the percent of AI/AN tribal lands with population losses after application of the privacy methods both remained close to 50 percent. The increase in Privacy Loss Budget between the two demonstration products did not seem to impact the overall percent of AI/AN tribal lands with gains and losses.

However, a difference can be seen when looking at the two white rows, the total number gained counts to the AI/AN tribal lands and total counts lost from AI/AN tribal lands. Although the percent of AI/AN tribal lands with gains and losses are somewhat balanced, the actual number of individuals gained and lost on AI/AN tribal lands is not balanced. With the exception of the demonstration product five (PLB 12.2) AI/AN Alone and In-Combination population, all three population categories for both demonstration products lost more population among all of the AI/AN tribal lands than was gained. This means that in both demonstration products, over 1,000 individuals that previously were on AI/AN tribal lands were no longer on AI/AN tribal lands. People who had identified as AI/AN that were counted on AI/AN tribal lands either were moved off of the tribal land or were no longer AI/AN due to the privacy protections. The percent of total tribal lands with gains and losses were balanced but the actual counts gained and lost were not with mostly reductions in counts after privacy methods were applied. There was some improvement with a higher PLB as would be expected, but the losses in counts on tribal lands is not insignificant.

The aggregate, or sum total, of the gains and losses provided an insight into the overall impacts on AI/AN tribal area populations from the planned 2020 Census privacy system. **Figure 7** further identifies the impacts on population losses on the top five highest total count losses from populations on AI/AN tribal lands. **Figure 7** shows the top five AI/AN tribal lands with the highest total population counts lost, AI/AN Alone counts lost, and AI/AN Alone and In-Combination population counts lost. The data illustrated is only for demonstration product five (PLB 12.2) to show the high losses of counts even with the higher level of accuracy applied.

**Figure 7** shows that regardless of population size, even in the demonstration product with the higher level of accuracy (PLB 12.2), AI/AN tribal lands with small or large populations can experience the highest levels of population losses compared to other AI/AN tribal lands. The percent of population loss varies based on the original population size, and the AI/AN tribal lands with smaller populations were disproportionately impacted with higher percent losses. Although the percent loss is smaller, larger AI/AN tribal lands are still potentially losing over 200 individuals from their populations, which is not an insignificant amount of people when, for example, federal funding is at stake.

**Figure 7. Top Highest Losses on any Tribal Area Type with Population Sizes in Demonstration Product 5 (PLB 12.2)**

Highest Population Count Loss			Highest AI/AN Alone Population Count Loss			Highest AI/AN Alone and In-Combination Population Count Loss		
Tribe and type of Tribal Area	Population group size	# (%)	Tribe and type of Tribal Area	Population group size	# (%)	Tribe and type of Tribal Area	Population group size	# (%)
United Houma Nation SDTSA	Above 100,000	-242 (-0.12%)	Apache Choctaw SDTSA	5,000 to 9,999	-99 (-6.68%)	Apache Choctaw SDTSA	5,000 to 9,999	-142 (-7.92%)
Chickasaw OTSA	Above 100,000	-216 (-0.07%)	Haliwa-Saponi SDTSA	5,000 to 9,999	-76 (-2.85%)	Chickasaw OTSA	Above 100,000	-123 (-0.30%)
Apache Choctaw SDTSA	5,000 to 9,999	-190 (-3.17%)	United Houma Nation SDTSA	Above 100,000	-71 (-0.90%)	Haliwa-Saponi SDTSA	5,000 to 9,999	-95 (-3.40%)
Echota Cherokee SDTSA	50,000 to 99,999	-164 (-0.31%)	Chickasaw OTSA	Above 100,000	-56 (-0.21%)	Echota Cherokee SDTSA	50,000 to 99,999	-93 (-2.59%)
Four Winds Cherokee SDTSA	25,000 to 49,999	-158 (-0.52%)	Waccamaw Siouan SDTSA	1,000 to 2,499	-54 (-4.30%)	Chickaloon ANVSA	10,000 to 24,999	-79 (-3.33%)

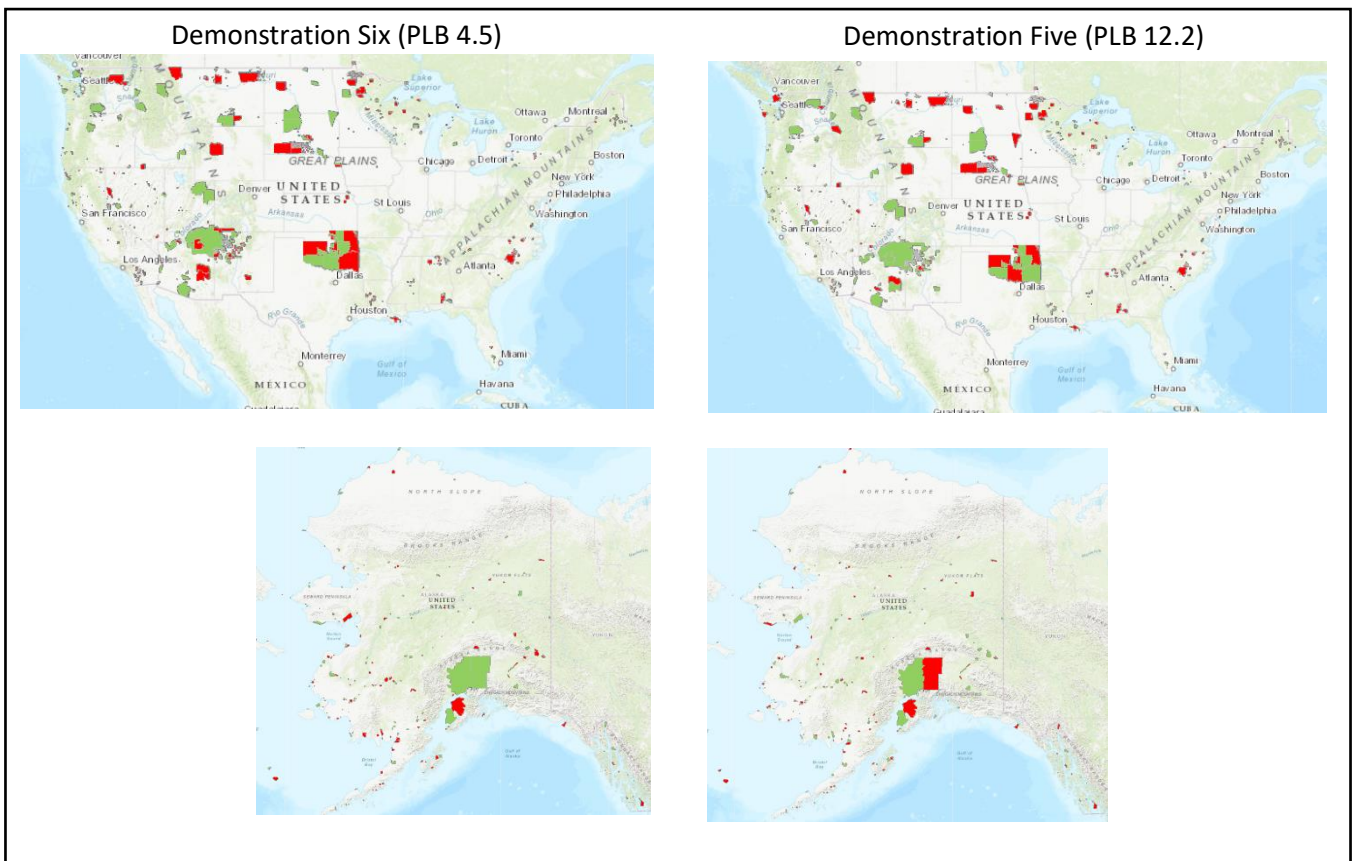
Every time the U.S. Census Bureau makes changes and processes the 2020 Census data through the 2020 Census Disclosure Avoidance System, Tribal Nations will be impacted in a random manner given the statistical nature of the privacy system. Once the U.S. Census Bureau make the final decisions on the structure of the privacy protections and processes the 2020 Census data, the resulting privacy protected data with the errors in it will be the official counts for at minimum the next ten years. **Figure 6** showed how the percent of AI/AN tribal lands with gains and losses remained relatively equal in distribution but the actual number of individuals gained and lost was not equal, and more counts were lost than gained. **Figure 7** showed the AI/AN tribal lands with the highest count losses in the demonstration product 5 (PLB 12.2), which has been described as being close to the final plans for the Disclosure Avoidance System.

**Figure 8** shows examples of the shift and randomness between Tribal Nations that experience a gain or a loss between demonstration products five and six using GIS maps of AI/AN tribal lands developed for this analysis. The shift between Tribal Nations that have population gains or losses is not predictable and can change any time data is processed through the algorithm. Regardless of how one Tribal Nation may have done in any demonstration product, every

demonstration product and the final Census tabulations under the current system remain somewhat of a gamble. Any Tribal Nation may be negatively impacted in the final 2020 Census dataset after the final privacy methods are applied.

The GIS map visualizations are interactive and free to use at the link below. They help users look at the impacts on specific Tribal Nations and the shifts between Tribal Nations with gains and losses through all six demonstration products. Visit our video tutorial on using the ArcGIS maps at <https://bit.ly/3eNfx43> and access the maps at <https://arcg.is/1fWG4u0>.

**Figure 8. Data Visualization of the Shift between AI/AN Tribal Lands with Population Gains and Losses in the April 2021 Demonstration Products**



The data from the two April 2021 demonstration products released by the U.S. Census Bureau provide an opportunity to assess what the potential impacts on AI/AN and tribal data might be in the actual 2020 Census data. This analysis covered some of the impacts on AI/AN tribal geography data from recent changes to the algorithm and the Privacy Loss Budget in the Disclosure Avoidance System applied to 2010 Census data. Removing the AI/AN tribal land population invariant does not appear to have helped the data and may have worsened the data for AI/AN tribal lands in the latest demonstration product. However, the increase in Privacy

Loss Budget does appear to have helped improve the AI/AN tribal geography data in some ways. However, the higher Privacy Loss Budget still saw some disproportionate impacts on different AI/AN tribal lands.

The U.S. Census Bureau will be making final decisions on the exact application of its Disclosure Avoidance System and the Privacy Loss Budget in early June 2021.<sup>16</sup> A final tribal consultation before the June decision is scheduled on Wednesday, May 19, 2021. Details on how to attend and to submit written comments for the tribal consultation are available at <https://bit.ly/3o7LQgO>. The deadline to submit final written comments for the current tribal consultation is May 28, 2021. The U.S. Census Bureau needs to hear from Tribal Nations on their priorities for uses of census data and the levels of accuracy in the data that are needed. Tribal Nations must decide if the price of privacy is worth the potential loss of accuracy in the 2020 Census data.

**Citation:** NCAI Policy Research Center (2021). *Impacts of the April 2021 Census Disclosure Avoidance System on Tribal Nations*. Washington DC: National Congress of American Indians, May 2021.

**Questions:** NCAI Policy Research Center – email: [research@ncai.org](mailto:research@ncai.org); website: <http://www.ncai.org/prc>

---

## Endnotes

<sup>1</sup> Jason Gauthier, History Staff. Title 13, U.S. Code - History - U.S. Census Bureau, [www.census.gov/history/www/reference/privacy\\_confidentiality/title\\_13\\_us\\_code.html](http://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html).

<sup>2</sup> Jason Gauthier, History Staff. Title 13, U.S. Code - History - U.S. Census Bureau, [www.census.gov/history/www/reference/privacy\\_confidentiality/title\\_13\\_us\\_code.html](http://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html).

<sup>3</sup> Bureau, US Census. “2020 Disclosure Avoidance System Updates.” The United States Census Bureau. [www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html](http://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html).

<sup>4</sup> Bureau, US Census. “2020 Disclosure Avoidance System Updates.” The United States Census Bureau. [www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html](http://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html).

<sup>5</sup> Bureau, US Census. “2020 Disclosure Avoidance System Updates.” The United States Census Bureau. [www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html](http://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html).

<sup>6</sup> U.S. Census Bureau. “2010 Demonstration Privacy-Protected Microdata Files 2021-04-28.” Developing the DAS: Demonstration Data and Progress Metrics, U.S. Census Bureau, <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20210428/2021-04-28-ppmf-factsheet.pdf>.

<sup>7</sup> U.S. Census Bureau. Developing the DAS: Demonstration Data and Progress Metrics: DAS Development Update 2021-04-28. <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html>.

<sup>8</sup> U.S. Census Bureau. Developing the DAS: Demonstration Data and Progress Metrics. <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html>.

---

<sup>9</sup> Van Riper D, Kugler T, Schroeder J. "IPUMS NHGIS Privacy-Protected 2010 Census Demonstration Data." IPUMS NHGIS, 1 May 2021, <https://nhgis.org/privacy-protected-demonstration-data#v20210428>.

<sup>10</sup> U.S. Census Bureau. "2010 Demonstration Privacy-Protected Microdata Files 2021-04-28." Developing the DAS: Demonstration Data and Progress Metrics, U.S. Census Bureau, <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20210428/2021-04-28-ppmf-factsheet.pdf>.

<sup>11</sup> Craige, Mary. "Implications to Montana of Differential Privacy for Census Bureau 2020 Data Dissemination." MT Department of Commerce. May 2021. [https://ftpaspen.msl.mt.gov/EventResources/20210413112706\\_19041.pdf](https://ftpaspen.msl.mt.gov/EventResources/20210413112706_19041.pdf).

<sup>12</sup> Craige, Mary. "Implications to Montana of Differential Privacy for Census Bureau 2020 Data Dissemination." MT Department of Commerce. May 2021. [https://ftpaspen.msl.mt.gov/EventResources/20210413112706\\_19041.pdf](https://ftpaspen.msl.mt.gov/EventResources/20210413112706_19041.pdf).

<sup>13</sup> U.S. Census Bureau. Developing the DAS: Demonstration Data and Progress Metrics: DAS Development Update 2021-04-28. <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html>.

<sup>14</sup> US Census Bureau. "2020 Disclosure Avoidance System Updates: 11/25/20 Invariants Set for 2020 Census Data Products" The United States Census Bureau, 25 Nov. 2020, [www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html](http://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html).

<sup>15</sup> US Census Bureau. "2020 Disclosure Avoidance System Updates: 11/25/20 Invariants Set for 2020 Census Data Products" The United States Census Bureau, 25 Nov. 2020, [www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html](http://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html).

<sup>16</sup> "The Road Ahead: Upcoming Disclosure Avoidance System Milestones." U.S. Census Bureau, 2010, <https://content.govdelivery.com/accounts/USCENSUS/bulletins/2c2e456>.

---

# THE IMPACT OF THE U.S. CENSUS DISCLOSURE AVOIDANCE SYSTEM ON REDISTRICTING AND VOTING RIGHTS ANALYSIS

---

**Christopher T. Kenny**  
Department of Government  
Harvard University  
Cambridge, MA  
christopherkenny@fas.harvard.edu

**Shiro Kuriwaki**  
Department of Government  
Harvard University  
Cambridge, MA  
kuriwaki@g.harvard.edu

**Cory McCartan**  
Department of Statistics  
Harvard University  
Cambridge, MA  
cmccartan@fas.harvard.edu

**Evan Rosenman**  
Harvard Data Science Initiative  
Harvard University  
Cambridge, MA  
erosenm@fas.harvard.edu

**Tyler Simko**  
Department of Government  
Harvard University  
Cambridge, MA  
tsimko@g.harvard.edu

**Kosuke Imai** \*  
Department of Government and Department of Statistics  
Harvard University  
Cambridge, MA  
imai@harvard.edu

May 28, 2021

## Abstract

The U.S. Census Bureau plans to protect the privacy of 2020 Census respondents through its Disclosure Avoidance System (DAS), which attempts to achieve differential privacy guarantees by adding noise to the Census microdata. By applying redistricting simulation and analysis methods to DAS-protected 2010 Census data, we find that the protected data are not of sufficient quality for redistricting purposes. We demonstrate that the injected noise makes it impossible for states to accurately comply with the *One Person, One Vote* principle. Our analysis finds that the DAS-protected data are biased against certain areas, depending on voter turnout and partisan and racial composition, and that these biases lead to large and unpredictable errors in the analysis of partisan and racial gerrymanders. Finally, we show that the DAS algorithm does not universally protect respondent privacy. Based on the names and addresses of registered voters, we are able to predict their race as accurately using the DAS-protected data as when using the 2010 Census data. Despite this, the DAS-protected data can still inaccurately estimate the number of majority-minority districts. We conclude with recommendations for how the Census Bureau should proceed with privacy protection for the 2020 Census.

**Keywords** Census · Redistricting · BISG · Differential privacy · TopDown algorithm · One Person One Vote

---

\*To whom correspondence should be addressed. We thank Ben Fifield and the ACLU for providing precinct-level state legislative assignments and election data for several states, and Bruce Willsie of L2, Inc for providing voterfiles.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Overview of Analysis</b>	<b>3</b>
2.1	Population Parity . . . . .	3
2.2	Partisan Effects . . . . .	4
2.3	Racial Effects . . . . .	4
2.4	Ecological Inference and Voting Rights Analysis . . . . .	5
<b>3</b>	<b>Summary of Findings</b>	<b>5</b>
<b>4</b>	<b>Population Parity in Redistricting</b>	<b>5</b>
4.1	Congressional Districts in Pennsylvania . . . . .	6
4.2	State Legislative Districts in Louisiana . . . . .	6
<b>5</b>	<b>Partisan Effects on Redistricting</b>	<b>7</b>
5.1	Partisan Patterns in DAS-induced Population Error . . . . .	7
5.2	Effects of Partisan Patterns on Aggregate Results . . . . .	8
<b>6</b>	<b>Racial Effects on Redistricting</b>	<b>9</b>
6.1	Racial Patterns in DAS-induced Population Error . . . . .	10
6.2	Effects of Racial Patterns on Aggregate and Precinct-level Results . . . . .	10
<b>7</b>	<b>Ecological Inference and Voting Rights Analysis</b>	<b>12</b>
7.1	Prediction of Individual Voter’s Race and Ethnicity . . . . .	13
7.2	Ecological Inference in the Voting Rights Analysis . . . . .	14
<b>8</b>	<b>Recommendations</b>	<b>16</b>

## 1 Introduction

In preparation for the official release of the 2020 Census data, the United States Census Bureau has built the Disclosure Avoidance System (DAS) to prevent Census respondents from being linked to specific people [1]. The DAS is based on differential privacy technology, which adds a certain amount of random noise to the raw Census counts. The decision to use differential privacy for the 2020 Census has been controversial, with many scholars voicing concerns about the negative impacts of noisy data on public policy and social science research, which critically rely upon the Census data [2, 3].

In this paper, we empirically evaluate the impact of the DAS on redistricting and voting rights analysis. Once released as part of the 2020 Census data later this year, states will use the P.L. 94-171 redistricting data to redraw their district boundaries of Congressional and other federal and local electoral offices. It is therefore of paramount importance to examine how the DAS affects redistricting analysis and the map-drawing process.

The Census Bureau has requested public feedback on the “fitness-for-use” of the P.L. 94-171 data by making available the Privacy-Protected Microdata Files (PPMFs) based on the application of the DAS to the 2010 Census redistricting data. The Census Bureau released two PPMFs at different levels of *privacy loss budget*,  $\epsilon$ , which controls the amount of noise. The DAS-12.2 data are based on a relatively high level of privacy loss budget ( $\epsilon = 12.2$ ) to achieve the accuracy targets at the expense of greater privacy loss, whereas the DAS-4.5 data use a lower privacy loss budget at the expense of worse accuracy ( $\epsilon = 4.5$ ). In addition, the Census Bureau post-processes the noisy data in order to ensure that the resulting public release data are self-consistent (e.g., no negative counts) and certain aggregate statistics such as state-level total population counts are accurate.



We examine the fitness-for-use of PPMFs through a variety of redistricting and voting rights analyses. In particular, we employ a set of recently developed simulation methods that can generate a large number of realistic redistricting maps under a set of legal and other relevant constraints, such as contiguity, compactness, population parity, and preservation of communities of interest and counties [4, 5, 6, 7, 8, 9, 10]. These simulation methods have been extensively used by expert witnesses in recent court cases on redistricting, including *Common Cause v. Lewis* (2020), *Rucho v. Common Cause* (2019), *Ohio A. Philip Randolph Institute v. Householder* (2020), *League of Women Voters of Michigan v. Benson* (2019), *League of Women Voters v. Pennsylvania* (2017), *Missouri State Conference of the NAACP v. Ferguson-Florissant School District* (2017), *Raleigh Wake Citizens Association v. Wake County Board of Elections* (2016), and *City of Greensboro v. Guilford County Board of Elections* (2015). These cases span all levels of government: local redistricting, state legislative redistricting, and congressional redistricting. We apply the simulation methods to the DAS-12.2 and DAS-4.5 data and compare the results with those obtained based on the 2010 Census data. This comparison reveals how the DAS affects the conclusions of redistricting analysis.

In addition, we examine the impact of DAS on the prediction accuracy of an individual voter’s race. Redistricting analysis for voting rights cases often necessitates such individualized prediction because most states’ voter lists do not include individual’s race. One prominent prediction method combines the Census block-level proportion of each race with a voter’s name and address [11, 12, 13]. This methodology played a key role in the most recent racial gerrymandering case, *NAACP, Spring Valley Branch et al. v. East Ramapo School District* (2020), in which the federal Court of Appeals for the Second Circuit upheld the district court’s ruling that the school board elections violated the Voting Rights Act. We reanalyze this case using the DAS data and compare the results with those based on the 2010 Census data.

## 2 Overview of Analysis

For the purposes of evaluating the impact of the new DAS on redistricting plan-drawing and analysis, we generated eight sets of redistricting datasets for simulation, described in Table 1. We create precinct-level datasets that have three versions of total population counts: the original 2010 Census, the DAS-12.2 data, and the DAS-4.5 data.

In our modal analysis, we simulate realistic district plans under the scenario that population counts are given by each of the three datasets. All simulations were conducted with the SMC redistricting sampler of [9], except for the Louisiana House of Representatives Districts for East Baton Rouge, which were conducted with a Merge-Split-type MCMC sampler similar to that of [5, 6]. Both of these sampling algorithms are implemented in the open-source software package `redist` [10]. All sampling diagnostics, including the number of effective samples, indicated accurate sampling and adequate sample diversity.

The DAS-12.2 data yield precinct population counts that are roughly 1.0% different from the original Census, and the DAS-4.5 data are about 1.9% different. For the average precinct, this amounts to a discrepancy of 18 people (for DAS-12.2) or 33 people (for DAS-4.5) moving across precinct boundaries. Therefore, our main simulation results should be thought of as a study of how such precinct-level differences propagate into noise at the district-level by exploring redistricting plans.

### 2.1 Population Parity

Perhaps the strongest constraint on modern redistricting is the requirement that districts be nearly equal in population. Deviations in population between districts have the effect of diluting the power of voters in larger-population districts. The importance of this principle stems from a series of Supreme Court cases in the 1960s, beginning with *Gray v. Sanders* (1963), in which the court held that political equality comes via a standard known as *One Person, One Vote*. As for acceptable deviations from population equality, *Wesberry v. Sanders* (1964) set the basic terms by holding that the Constitution requires that “as nearly as is practicable, one person’s vote in a congressional election is to be worth as much as another’s.” Even minute differences in population parity across congressional districts must be justified, even when smaller than the expected error in decennial Census figures (*Karcher v. Daggett* 1983). For state legislative districts, *Reynolds v. Sims* (1964) held that they must be drawn to near population equality. However, subsequent rulings stated that states may allow for small population deviations when seeking other legitimate interests (*Mahan v. Howell* 1972; *Gaffney v. Cummings* 1973).

When measuring population equality, states must rely on Census data, which was viewed as the most reliable source of population figures (*Kirkpatrick v. Preisler* 1969). We therefore empirically examine how the DAS affects the ability to draw redistricting maps that adhere to this equal population principle. We simulate

State	Office	Districts	Precincts	Total simulated plans
Pennsylvania	U.S. House	18	9,256	30k
Louisiana	State Senate	39	3,668	60k
Louisiana*	State House	15	361	1,700k
North Carolina	U.S. House	13	2,692	30k
South Carolina	U.S. House	7	2,122	30k
South Carolina	State House	124	2,122	30k
Mississippi <sup>§</sup>	State Senate	9	310	30k
New York <sup>†</sup>	School Board	9	1,207	10k

**Table 1:** States and districts studied. We compared the Census 2010, DAS-12.2, and DAS-4.5 datasets in six states and three levels of elections.

\*Examines the Baton Rouge area.

<sup>§</sup>Examines District 22 and its 8 adjacent districts.

<sup>†</sup>Examines the East Ramapo school district, using Census blocks instead of voting precincts.

realistic maps for Pennsylvania Congressional districts and Louisiana State Senate districts based on the DAS-4.5 and DAS-12.2 data under various levels of population parity. We then examine the degree to which the resulting maps satisfy the same population parity criteria using the 2010 Census data.

## 2.2 Partisan Effects

If changes in reported population in precincts affect the districts in which they are assigned to, this has implications for which parties win those districts. While a change in population counts of about 1 percent may seem small, differences in vote counts of that magnitude can reverse some election outcomes. Across the five U.S. House elections during 2012 – 2020, 25 races were decided by a margin of less than a percentage point between the Republican and Democratic party’s vote shares. And 228 state legislative races were decided by less than a percentage point from 2012–2016.

Partisan implications also raise the concern of gerrymandering, where political parties draw district boundaries to systematically favor their own voters. Many uses of redistricting simulation in redistricting litigation have been over partisan gerrymanders, including *Common Cause v. Lewis*, *Rucho v. Common Cause*, *Ohio A. Philip Randolph Institute v. Householder*, *League of Women Voters of Michigan v. Benson*, and *League of Women Voters v. Pennsylvania*. To evaluate the impact of the DAS on the analysis of potential partisan gerrymanders, we simulate 120,000 redistricting plans across the states of Pennsylvania, North Carolina, and South Carolina, and compare the partisan attributes of the simulated plans from the three data sources. We also analyze voting-related patterns in DAS-induced population count error at the precinct level, and connect these patterns to the statewide findings from the simulations.

## 2.3 Racial Effects

The Voting Rights Act of 1965, its subsequent amendments, and a series of Supreme Court cases all center race as an important feature of redistricting. A large number of these cases focus on the creation of majority-minority districts (MMDs) (e.g. *Thornburg v. Gingles* 1986, *Shaw v. Reno* 1993, *Miller v. Johnson* 1995, *Shelby County v. Holder* 2013). First, we analyze whether the DAS data systematically undercounts or overcounts certain areas across racial lines. In doing so, we focus attention on the potential consequences of the decision to target accuracy to the majority racial group in a given area [14]. We explore patterns with racial diversity in four states (Pennsylvania, Louisiana, North Carolina, South Carolina).

We also explicitly explore how DAS data can influence the creation of MMDs. To do so, we empirically examine how using the DAS data to create MMDs differs from the same process undertaken using the 2010 Census data. We simulate nearly two million maps in the Louisiana State House and examine the degree to which maps generated using the Census and DAS data lead to different results at the district and precinct levels.

## 2.4 Ecological Inference and Voting Rights Analysis

Social scientists have developed methods to predict the race and ethnicity of individual voters using Census data. Since *Gingles*, voting rights cases have required evidence that an individual’s race is highly correlated with candidate choice. Statistical methods must therefore estimate this individual quantity from aggregate election results and aggregate demographic statistics [15, 16]. A key input to these methods is accurate racial information on voters. To produce this data, recent litigation has used Bayesian Improved Surname Geocoding (BISG) to impute race and ethnicity into a voter file [11, 12, 13]. This methodology is often used to improve classification of the degree of racially polarized voting and racial segregation.

To understand how DAS data influence these analyses, we look at the effect of DAS data on BISG accuracy across several states where race is recorded on the voter file. We then re-examine a recent voting rights case on a school board election in New York using the DAS-12.2 data and compare results to using the Census 2010 data.

## 3 Summary of Findings

Compared to the original Census 2010 data, we find that the DAS-protected data:

- **Prevent map drawers from creating districts of equal population, according to current statutory and judicial standards.** Actual deviations from equal population will generally be several times larger than as reported under the DAS data. The magnitude of this problem increases for smaller districts such as state legislative districts and school boards.
- **Transfer population from low-turnout, mixed-party areas to high-turnout, single-party areas.** This differential bias leads to different district boundaries, which in turn implies significant and unpredictable differences in election results. The discrepancy also degrades the ability of analysts to reliably identify partisan gerrymanders.
- **Transfer population from racially mixed areas to racially segregated areas.** This bias effectively means racially heterogeneous areas are under-counted. The degree of racial segregation can therefore be over-estimated, which can lead to a change in the number of majority-minority districts. It also creates significant precinct-level variability, which adds substantial unpredictability to whether or not a minority voter is included in a majority-minority district.
- **Alter individual-level race predictions constructed from voter names and addresses.** This leads to fewer estimated minority voters and majority-minority districts in a re-analysis of a recent Voting Rights Act case, *NAACP v. East Ramapo School District*. At a statewide level, however, the DAS data does not curb the ability of algorithms to identify the race of voters from names and addresses. Therefore, this casts doubt on the universal privacy protection guarantee of DAS data.

The subsequent sections deal with these findings and their accompanying methods and data in more detail.

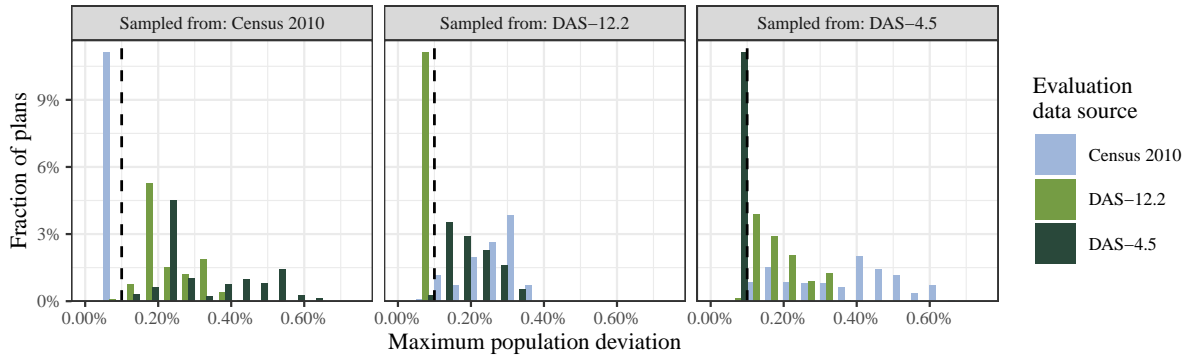
## 4 Population Parity in Redistricting

Deviation from population parity across  $n_d$  districts is generally defined as

$$\text{deviation from parity} = \max_{1 \leq k \leq n_d} \frac{|P_k - \bar{P}|}{\bar{P}},$$

where  $P_k$  denotes the population of district  $k$  and  $\bar{P}$  denotes the target district population. In other words, we track the percent difference in the district population  $P_k$  from the average district size  $\bar{P}$ , and report the maximum deviation. Our redistricting simulations generate plans that do not exceed a user-specified tolerance. After generating these plans, we then re-evaluate the deviation from parity using the precinct populations from the three data sources.

We find that the noise introduced by the DAS prevents the drawing of equal-population maps with commonly-used population deviation thresholds. Because only one dataset will be available in practice, redistricting practitioners who attempt to create equal-population districts with DAS data should expect the actual



**Figure 1:** Maximum deviation from population parity among Pennsylvania redistricting plans simulated from the three data sources. All plans were sampled with a population constraint of 0.1 percent, corresponding to the deviation measured from the Census 2010 precinct data, and marked with the dashed line. Deviation from parity was then evaluated using the three versions of population data.

deviation from parity to be significantly larger than what they can observe in their data. This problem is more acute in state legislative districts, where there are more districts and each district is comprised of fewer precincts.

#### 4.1 Congressional Districts in Pennsylvania

Figure 1 shows the maximum deviation from population parity for the 30,000 simulated redistricting plans in Pennsylvania, when evaluated according to the three different data sources.<sup>2</sup> Consistently, plans generated under one set of population data and drawn to have a maximum deviation of no more than 0.1% had much larger deviations when measured under a different set of population data. For example, of the 10,000 maps simulated using the DAS-12.2 data, 9,915 exceeded the maximum population deviation threshold, according to the Census 2010 data. While nearly every plan failed to meet the population deviation threshold, the exact amount of error varied significantly across the simulation set. As a result, redistricting practitioners who attempt to create equal-population districts according to similar thresholds can expect the actual deviation from parity to be significantly larger but of unknown magnitude.

#### 4.2 State Legislative Districts in Louisiana

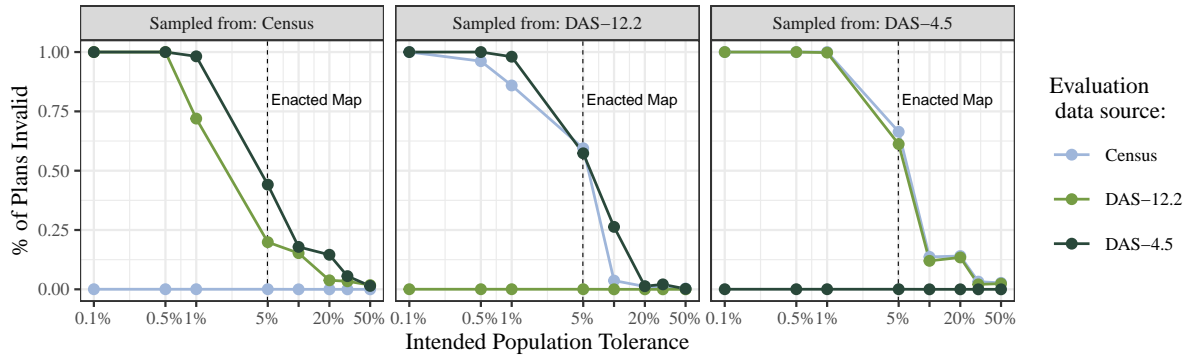
We expect smaller districts such as state legislative districts to be more prone to discrepancies in population parity. For example, the average Louisiana Congressional district comprises about 600 precincts, but a State Senate district comprises about 90 and a State House district only 35. Therefore, deviations due to DAS are more likely to result in larger percent deviations from the average. To test this, we compared 60,000 Louisiana State Senate plans generated from the three data sources and population parity constraints ranging from 0.1% to 50%, measuring the plans' population deviation against the three different data sources.<sup>3</sup> Figure 2 plots the results of this comparison.

As expected, we see complete acceptance for plans measured with the dataset from which they were generated. However, plans generated under one dataset can be invalid under another. Specifically, plans generated under DAS data can be very likely to be invalid when evaluated using the true Census data. The rate of invalid plans grows as the tolerance becomes more precise.

Also noteworthy is the fact that even at the population parity tolerances as generous as 1.0%, all generated plans are invalid in some cases. Compared to Pennsylvania, with a parity tolerance of 0.1%, this is as a result

<sup>2</sup>10,000 plans were simulated from each data source, with every plan satisfying a 0.1% population parity constraint. The simulation algorithm also ensured that no more than 17 counties were split across the entire state, reflecting the requirement in Pennsylvania that district boundaries align with the boundaries of political subdivisions to the greatest extent possible.

<sup>3</sup>2,500 plans were simulated for each data source/population parity pair.



*Figure 2: Fraction of Louisiana State Senate plans simulated under one data source which are invalid when measured under another. The dashed line shows the parity of the enacted 2010 map.*

of the smaller district sizes in the Louisiana State Senate—the DAS-added noise is relatively larger at smaller scales.

## 5 Partisan Effects on Redistricting

To analyze the partisan implications of a redistricting plan using a set of simulated redistricting plans, practitioners generate hypothetical district-level election results for the simulated plans and for the plan to be analyzed. Plans which are partisan gerrymanders stand out from the simulated ensemble as yielding more seats for one party over the other.

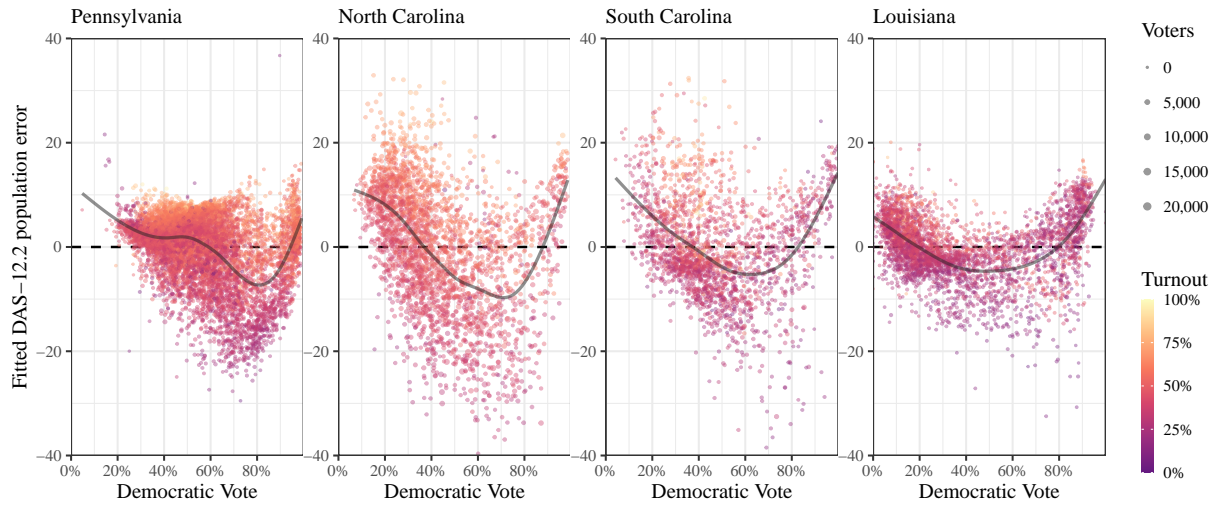
In computing a party’s expected vote share for each congressional district, we use data from statewide elections to avoid the variation in uncontested races and any incumbency effects in U.S. House races. In Pennsylvania, we use the two party vote share averaged across all statewide and Presidential races, 2004–2008, and adjust to match 2008 turnout levels. In South Carolina we use the 2018 gubernatorial election, in North Carolina we use the 2012 gubernatorial election, and in Louisiana we use the 2019 Secretary of State election.

### 5.1 Partisan Patterns in DAS-induced Population Error

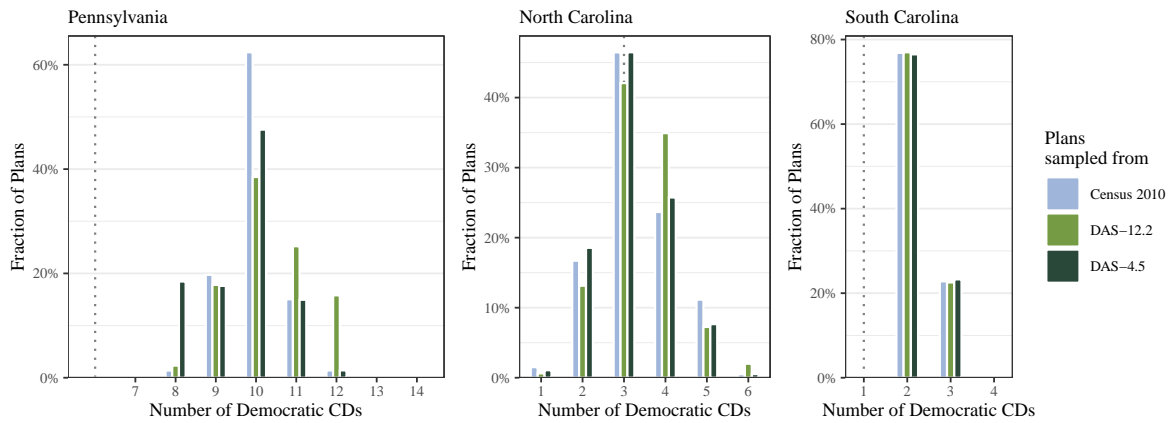
We first examine the electoral correlates of population change induced by the DAS. By the nature of the noise injection of the DAS, there is significant variation in the population error, even among similar precincts, and as a result it is difficult to discern systematic patterns by observation alone. Consequently, we fit a generalized additive model (GAM) to the precinct-level population errors to understand the degree to which different factors influence these errors, on average. The GAM regresses the difference in precinct population between the DAS-12.2 and the Census data on a tensor product cubic regression smooth of precinct turnout, two-way Democratic vote, and log population density, and thin-plate regression splines of the fraction of voters who are White and the racial Herfindahl-Hirschman index [17, 18]. We fit the GAM on precincts in Pennsylvania, North Carolina, South Carolina, and Louisiana. The model explained about 9–12 percent of the overall variance in population errors.

Figure 3 plots the fitted values from this model against Democratic vote share for each of the four states. Perhaps unexpectedly, several consistent patterns emerge. First, higher-turnout precincts are on average assigned more population under the DAS than they should otherwise have, according to the 2010 Census. Second, moderately Democratic precincts are on average assigned less population under the DAS. These effects are on the order of 5–15 voters per precinct, on average, though some are larger.<sup>4</sup> Aggregated across the hundreds of precincts that comprise the average district, however, the errors may become substantial. In Pennsylvania’s 2nd and 3rd Congressional Districts, for example, which cover Philadelphia County and are majority-minority, the accumulated population error in each district is on average 3,000 voters across the set of simulated plans.

<sup>4</sup>Not shown is the equivalent figure for the DAS-4.5 data, which displayed an identical pattern but with roughly double the magnitude of fitted error.



**Figure 3:** Model-smoothed error in precinct populations by Democratic two-party vote share, with color indicating turnout. A GAM smooth is overlaid to show the mean error by Democratic share.

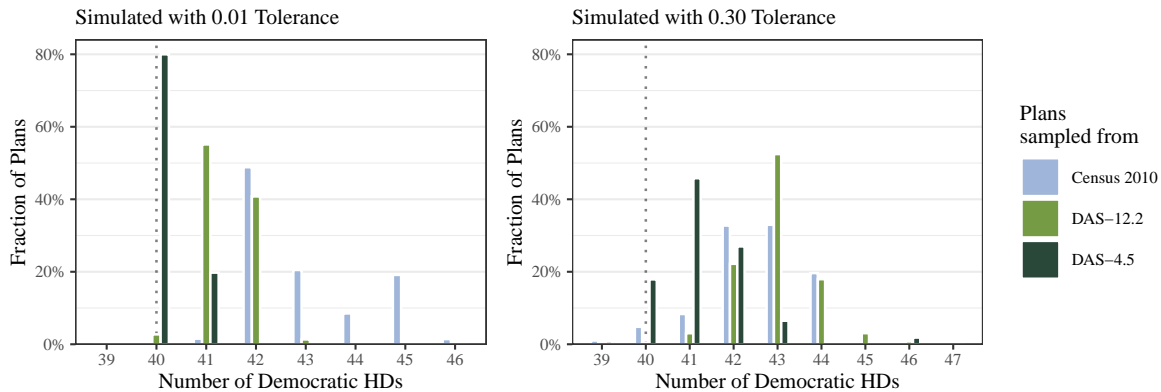


**Figure 4:** Distribution of Democratic-majority congressional districts, by data source and state. The vertical dashed lines indicated the number of Democratic-majority seats under the plans enacted by the state legislatures.

Some of these partisan effects may be explained by racial patterns, as shown in Figure 6 and discussed below in Section 6. It is difficult to know exactly these partisan and racial biases arise without more detail on the DAS post-processing system and parameters. Regardless, the presence of differential bias in the precinct populations according to partisanship and turnout is concerning. These precinct-level biases may aggregate in unexpected ways, leading to potentially large unknown biases in statewide analyses, as we discuss next.

### 5.2 Effects of Partisan Patterns on Aggregate Results

The spatial distribution of these types of precincts, and the details of the DAS post-processing, critically determine the overall effect once these precincts are aggregated into larger districts. Given the results of Figure 3, we would expect that aggregation to districts may not cancel out DAS-induced noise entirely. Indeed, for the 44 congressional districts in the four states we examine, the average district’s population changes by 292 people (or 1%) by DAS-12.2 data, but in three Pennsylvania congressional districts in and around Philadelphia, the population changes by 1,311 people on average. Two congressional races in these



**Figure 5:** Distribution of Democratic-majority South Carolina State House districts.

four states have been decided by less than a percentage point during 2012-2020: NC-07 in 2012 and NC-09 in 2018.

We find that the DAS leads to unpredictable differences in the distribution state-level party outcomes under the three data sources. Figure 4 compares the distribution of the number of congressional districts in which the Democratic Party’s candidate wins over 50% of the two-party vote.<sup>5</sup> In Pennsylvania and North Carolina, plans simulated with DAS-12.2 tend to favor the Democratic party more than plans simulated with DAS-4.5 or the original Census. The implied number of Democratic seats in the *enacted* plans, shown in the dotted line, tend to be on the lower end of the simulated reference distribution, although our simulations here do not impose constraints required by the Voting Rights Act.

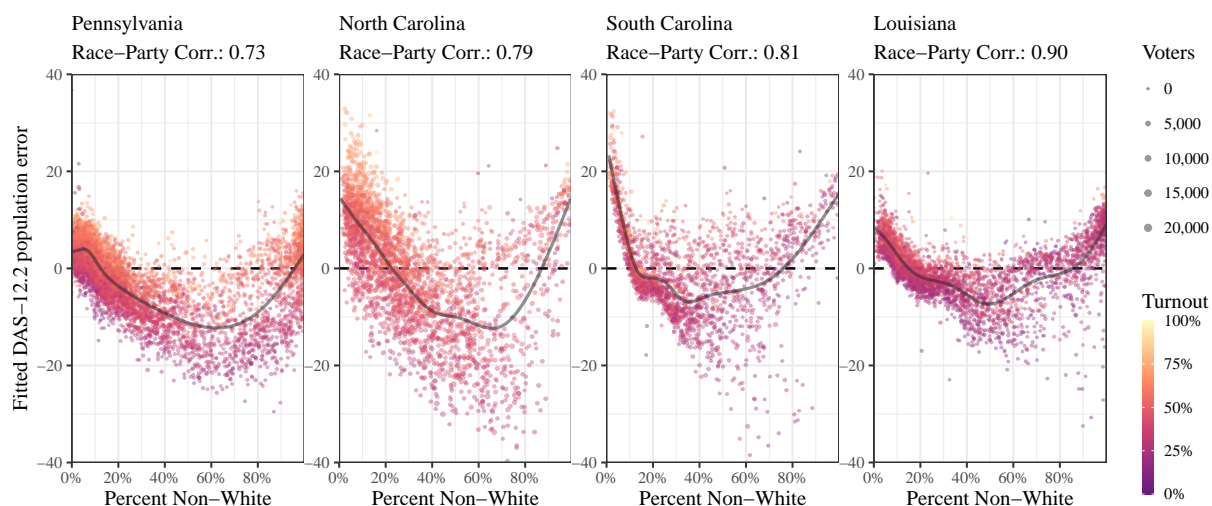
Interestingly, with congressional districts, the DAS-4.5 data tend to produce a distribution of Democratic seats closer to the 2010 Census, even though it is noisier than DAS-12.2 on average. We caution that the number of congressional districts with majority Democratic vote is a coarse measure and can mask more subtle differences. For example, in South Carolina, the overall distribution of Democratic seats does not differ, but this may mask differences captured by other continuous metrics like mean-median difference in voteshares.

Differences between data sources are likely more stark for state legislative districts, which are composed from fewer precincts than the congressional districts. In Figure 5 we show simulations from the state legislative districts in South Carolina. We show two simulations with different tolerances for deviations from population parity.

Once again, there are significant differences in the distribution of Democratic seats across the three data sources, but the pattern in location and scale changes are not monotonic with the level of noise. Notably, at a 1% population parity constraint, the enacted legislative map is an outlier under the Census 2010 and DAS-12.2 simulations, but is the modal outcome under the DAS-4.5 data. A discrepancy of this magnitude could change the factual findings regarding the presence or absence of a partisan gerrymander in redistricting litigation.

## 6 Racial Effects on Redistricting

We also investigate the potential impact of privacy-protected data on the role of race in redistricting. We begin by the analysis of racial patterns in the population errors induced by the DAS. We then examine how those racial biases affect redistricting outcomes.



**Figure 6:** Model-smoothed error in precinct populations by the minority fraction of voters, with color indicating turnout. A GAM smooth is overlaid to show the mean error by minority share.

### 6.1 Racial Patterns in DAS-induced Population Error

In the previous section, we demonstrated that the population error introduced by the DAS procedure overcounts the most homogeneous Republican and Democratic precincts in high-turnout areas and undercounts heterogeneous, low-turnout areas. Race is highly correlated with partisanship in American politics, and we find that the same pattern of differential error by race and turnout levels holds for race as well as partisanship. Figure 6 shows this pattern across the states we have analyzed so far (PA, LA, NC, and SC). The results imply that in terms of population error, mixed White/nonwhite precincts lose the most population relative to more homogeneous precincts. Figure 7 more clearly shows this pattern with homogeneous precincts. We plot the error against the Herfindahl-Hirschman Index and find that the fitted error in estimated population steeply declines as the precinct becomes more racially diverse.

These patterns are likely partially explained by the adopted DAS targets [14], which prioritize accuracy for the largest racial group in a given area. By doing so, the DAS procedure appears to undercount heterogeneous areas where the population differences between racial groups are relatively small. As precincts are the building blocks of political districts, our results demonstrate that precincts that are heterogeneous along racial and partisan lines would lose electoral power under the DAS. In aggregate, the movement of population from heterogeneous to homogeneous precincts would tend to increase the apparent spatial segregation by race.

### 6.2 Effects of Racial Patterns on Aggregate and Precinct-level Results

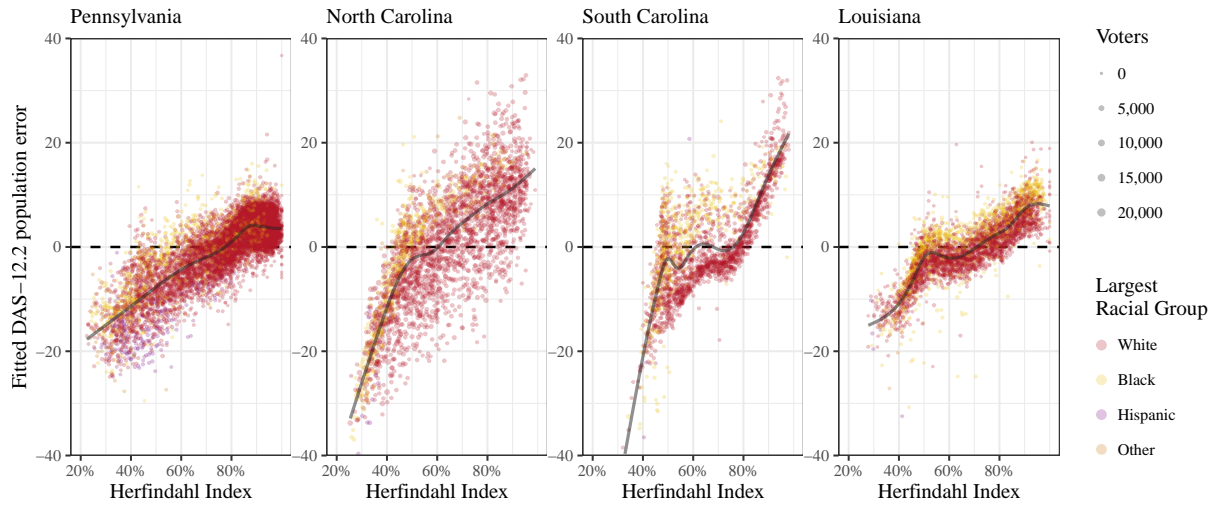
As with the partisan patterns of DAS population bias, the racial patterns of bias may not necessarily cancel upon aggregation. To evaluate the impact of these biases, we compare the distribution of the number of majority-minority districts (MMDs) across the simulations from the three data sources. MMDs are a primary focus in voting rights litigation and the analysis of race in redistricting.

Figure 8 shows the effects of the DAS on the number of MMDs in the South Carolina state House and Mississippi state Senate.<sup>6</sup> Ten thousand plans simulated from both 2010 Census and DAS-12.2 data were evaluated for MMDs under both data sources. There are two types of discrepancies. Not visible in the figure is the fact that while generally the DAS and 2010 Census data agree on the presence of an MMD given a set of simulated plans, the DAS data slightly but systematically understate the number of such districts in

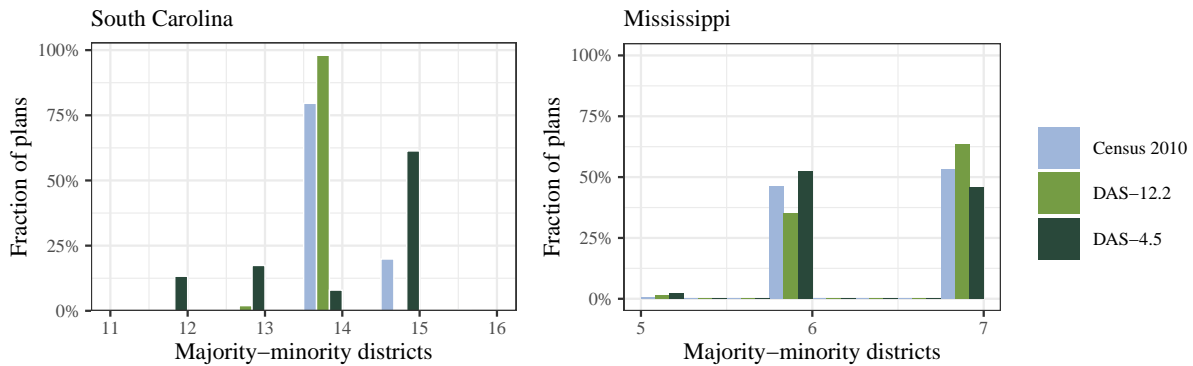
<sup>5</sup>Simulations in North Carolina and South Carolina shown here satisfy a 1% population parity constraint, and ensure that no more than 12 and 6 counties, respectively, are split in each plan. Data for North Carolina was obtained from the North Carolina General Assembly Redistricting Archives.

<sup>6</sup>The Mississippi plans were generated to satisfy a 5.0% population parity constraint, reflecting the 4.98% population parity deviation of the currently enacted plan. Data for Mississippi was obtained from the Mississippi Automated Resource Information System.





**Figure 7:** Model-smoothed error in precinct populations by the Herfindahl-Hirschman Index. A Herfindahl-Hirschman Index of 100 percent indicates that the precinct is comprised of only one racial group.

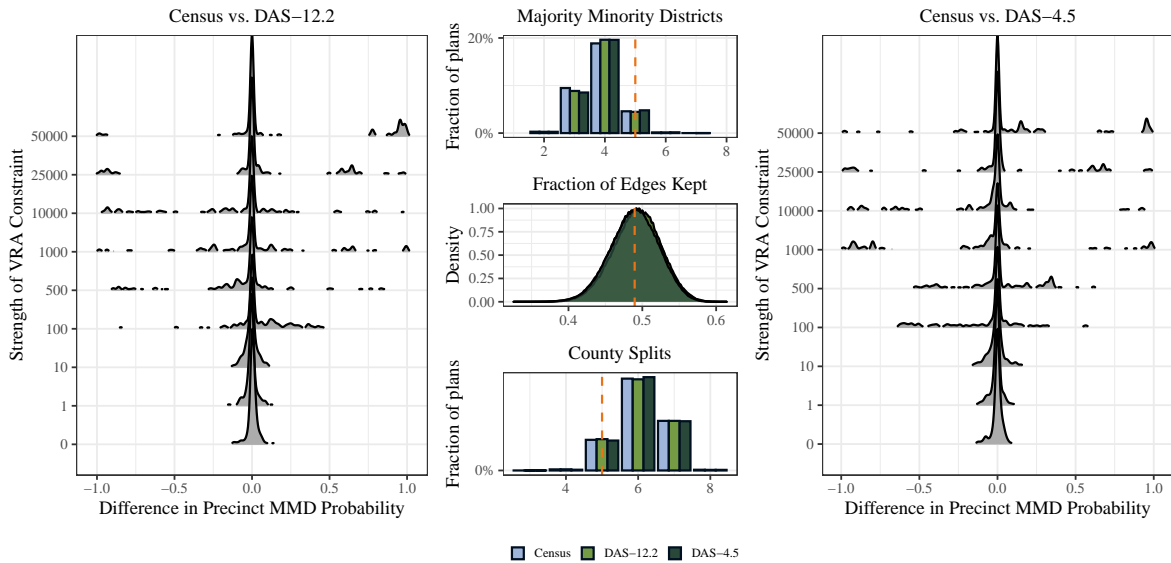


**Figure 8:** Distribution of majority-minority districts in South Carolina and Mississippi, by simulation data source.

South Carolina and overstate the number of such districts in Mississippi. For example, in South Carolina, among the 1,986 plans simulated from 2010 Census data that had 15 MMDs, 4.7% had only 14 MMDs when evaluated with DAS-12.2 data.

What is more concerning is that the overall distribution of the number of MMDs is significantly different across data sources. In Mississippi, the DAS-12.2 data generates far fewer plans with 6 MMDs compared to the 2010 Census data. In South Carolina, meanwhile, there are no simulated plans with 15 MMDs under DAS-12.2 data, but such plans make up nearly 20% of the 2010 Census-based simulations. As a result, a legislature-adopted plan drawn with 15 MMDs according to DAS-protected data could be improperly classified as an extreme outlier and might even be struck down as a racial gerrymander.

If these differences between DAS-based and 2010 Census-based summary statistics were of predictable magnitude, it might be possible for states or analysis to adjust to the additional noise. However, as with the partisan effects, we find that the DAS-induced distortions are not necessarily consistent across states. Our primary case for this purpose is Louisiana’s East Baton Rouge Parish and the surrounding area. We chose this area because the city of Baton Rouge includes a large Black population represented by multiple MMDs in the state’s lower and upper houses. From the 15 lower house districts in this area (each with approximately 40,000 population) comprising 361 precincts, we simulate 500,000 plans under each of the three data sources. We simulate each plan with a maximum 5% population parity constraint to match the enacted map. For each



**Figure 9:** The center column shows district-level comparisons between 500,000 plans generated under 2010 Census data, DAS-4.5 data, and DAS-12.2 data. Few aggregate-level differences are seen across three commonly used metrics—the number of majority-minority districts, the number of parish (county) splits, and the compactness of the districts. However, the left and right columns show that precinct-level assignments can differ substantially between the 2010 Census and DAS data. Here, the calculated probability of being assigned to a majority-minority district can be much higher or lower for individual precincts, and these differences grow as a constraint encouraging the formation of MMDs is strengthened.

of these plans, we measure three commonly used metrics in redistricting—the number of resulting MMDs, the number of parish splits, and the compactness of the plan.

The middle column of Figure 9 finds few district-level differences between plans generated using 2010 Census data versus DAS data. Plans generated under all three datasets have essentially identical distributions of MMDs, parish splits, and compactness.

However, these aggregate distributions mask the variability around which individual precincts are included in majority minority districts. In the left and right columns of Figure 9, we show the results of 10,000 simulations of the Merge-Split-type MCMC sampler with various levels of a Voting Rights Act (VRA) constraint. This constraint, which we did not apply in the previous sections, encourages the formation of majority-minority districts. We then calculate the probability that each precinct is assigned to a majority-minority district (as defined by Black population). Finally, we calculate the difference between these probabilities for the Census versus DAS-12.2 and Census versus DAS-4.5.

With no VRA constraint, each precinct has similar probabilities of being in a MMD, regardless of the dataset used. However, as the strength of this constraint increases (making the algorithm search for MMDs more aggressively), we see that the noise introduced to the DAS data systematically alters the district membership of individual precincts. A precinct with a value of 1 or  $-1$  in the left and right columns of Figure 9 indicates that those precincts are never in a MMD under one dataset but are always in a MMD when the same mapmaking process is done with a different dataset.

## 7 Ecological Inference and Voting Rights Analysis

Inferring the racial and ethnic composition of potential voters and their candidate choice is a key element of voting rights analysis in redistricting. Recent court cases have relied on Bayesian Improved Surname Geocoding (BISG) to predict the race and ethnicity of individual voters in a voter file [11, 12, 13]. This

methodology combines the names and addresses of registered voters with block-level racial composition data from the Census.

We first examine how the accuracy of prediction changes between the DAS and original Census data. Since this is exactly the type of analysis from which the DAS is supposed to protect individual Census respondents, we expect the prediction accuracy to dramatically decline when using the DAS-protected data. We then revisit the most recent court case about the East Ramapo school board election and investigate whether this change in racial prediction alters the conclusions of the racial redistricting analysis.

### 7.1 Prediction of Individual Voter’s Race and Ethnicity

We first compare the accuracy of predicting individual voters’ race and ethnicity using the original 2010 Census data, the DAS-12.2 data, and the DAS-4.5 data. To obtain the benchmark, we use the North Carolina voter file obtained in February 2021.<sup>7</sup> In several southern states including North Carolina,<sup>8</sup> the voter files contain the self-reported race of each registered voter. This information can then be used to assess the accuracy of the BISG prediction methodology.

Our approach follows [19]. We denote by  $E_i$  the ethnicity of voter  $i$ ,  $N_i$  as the surname of voter  $i$ , and  $G_i$  as the geography in which voter  $i$  resides. For each choice of ethnicity  $e \in \mathcal{E} = \{\text{White, Black, Hispanic, Asian, Other}\}$ , Bayes’ rule implies

$$P(E_i = e \mid N_i = n, G_i = g) = \frac{\Pr(N_i = n \mid E_i = e) \Pr(E_i = e \mid G_i = g)}{\sum_{e' \in \mathcal{E}} \Pr(N_i = n \mid E_i = e') \Pr(E_i = e' \mid G_i = g)},$$

where we have assumed the conditional independence between the surname of a voter and their geolocation within each racial category, i.e.,  $N_i \perp\!\!\!\perp G_i \mid E_i$ .

In the presence of multiple names—e.g. first name  $f$ , middle name  $m$ , and surname  $s$ —we make the further conditional independence assumption [20]

$$\Pr(N_i = \{f, m, s\} \mid E_i = e) = \Pr(F_i = f \mid E_i = e) \Pr(M_i = m \mid E_i = e) \Pr(S_i = s \mid E_i = e),$$

where  $F_i$ ,  $M_i$ , and  $S_i$  represent individual  $i$ ’s first, middle, and surnames respectively.

We compare estimates by changing the data source from which the geographic prior,  $\Pr(E_i = e \mid G_i = g)$ , is estimated, from the 2010 Census to each of the two DAS datasets. Estimates of the other race prediction probabilities are obtained by merging three sources: the 2010 Census surname list [21], the Spanish surname list from the Census, and the voter files from six states in the U.S. South, where state governments collect racial and ethnic data about registered voters for Voting Rights Act compliance. The middle and first name probabilities are derived exclusively from the voter files.

We evaluate the accuracy of the BISG methodology on approximately 5.8 million registered voters included in the North Carolina February 2021 voter file. Among them, approximately 70% are White and 22.5% are Black, with smaller contingents of Hispanic (3.4%), Asian (1.5%), and Other (2.4%) voters.

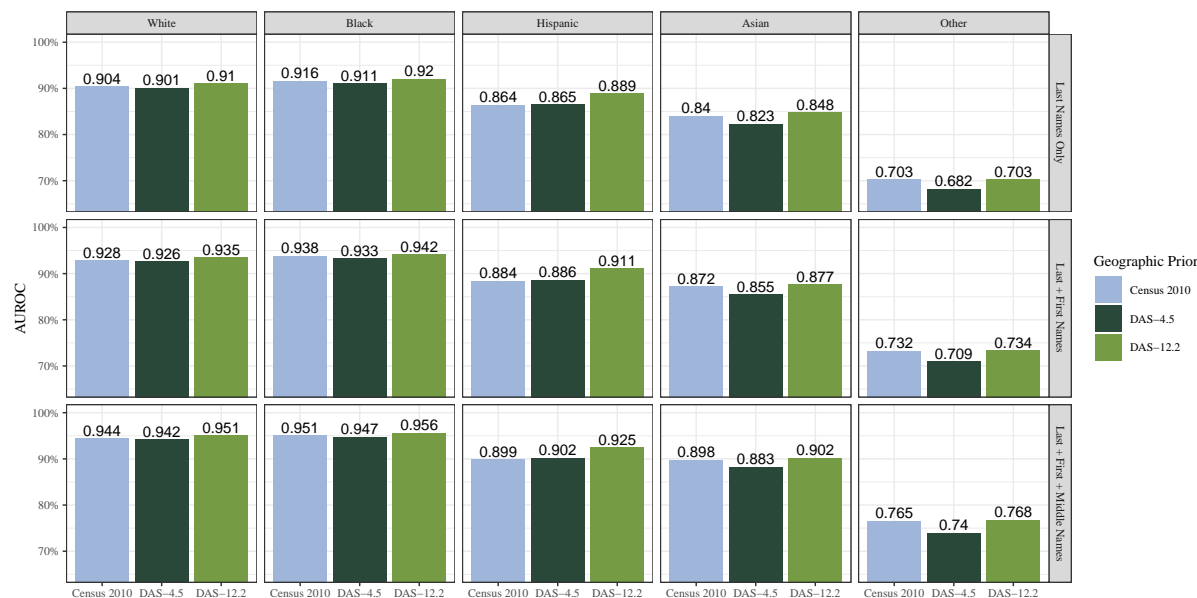
Figure 10 summarizes the accuracy of the race prediction with the area under the Receiver Operating Characteristic curve (AUROC). The AUROC ranges from 0 (perfect misclassification) to 1 (perfect classification). Across all racial and ethnic groups except Hispanics, we find the same surprising pattern: relative to the 2010 Census data, the DAS-12.2 data yield a small improvement in prediction performance while the DAS-4.5 data give a slight degradation. Among Hispanics, both forms of DAS-protected data result in slightly improved predictions over the original Census data.

The strong performance of the DAS-12.2 data in this setting is counter-intuitive. It is possible that the noise added to the underlying data has somehow mirrored the true patterns of population shift from 2010 to 2021; or that this noise makes the DAS-12.2 data more reflective of the voter population relative to the voting-age population. Additionally, the DAS may degrade or attenuate individual probabilities without having a significant impact on the overall ability to classify, something that AUROC is not designed to measure [22].

Results are substantively similar if we consider the classification error, under the heuristic that we assign each individual to the ethnic group with the highest posterior probability. Using the true census data to establish

<sup>7</sup>We obtain the voter files used in this paper through L2, Inc., which is a leading national nonpartisan firm that supplies voter data and related technology.

<sup>8</sup>The other states are Alabama, Florida, Georgia, Louisiana, and South Carolina



**Figure 10:** Area under the Receiver Operating Characteristic Curve (AUROC) percentage values for the prediction of individual voter's race and ethnicity using North Carolina voter file. Bars represent AUROC with geographic priors given by each of three datasets: 2010 Census, DAS-4.5, and DAS-12.2.

geographic priors, we achieve posterior misclassification rates of 15.1%, 12.1%, and 10.0% when using the last name; last name and first name; and middle names for prediction, respectively. The analogous misclassification rates are slightly higher for the DAS-4.5 priors—15.6%, 12.5%, and 10.3%—but the same or slightly lower for the DAS-12.2 priors: 15.1%, 12.0%, and 9.9%.

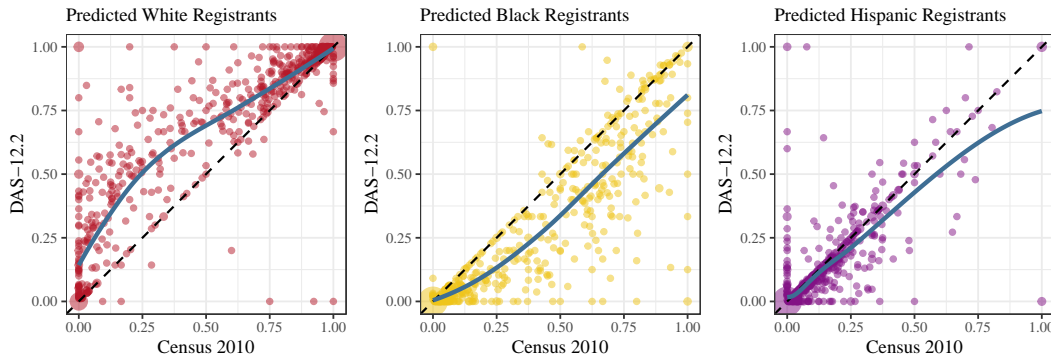
Our analysis shows that across three main racial and ethnic groups, the predictions based on the DAS data appear to be as accurate as those based on the 2010 Census data. The finding suggests that the DAS data may not provide universal privacy protection.

## 7.2 Ecological Inference in the Voting Rights Analysis

The BISG methodology played a central role in the most recent court case regarding Section 2 of the Voting Rights Act, *NAACP of Spring Valley v. East Ramapo Central School District* (2020). The East Ramapo Central School District (ERCSD) nine-member school board was elected using at-large elections. This often led to an all White school board, despite 35% of the voter eligible population being Black or Hispanic. Yet, within the district, nearly all White school children attend private yeshivas, whereas nearly all Black and Hispanic children attend the ERCSD public schools. As a result of this case, the district moved to a ward system.

We re-examine the remedy of this case, focusing on effective majority-minority districts (MMDs) based on a voter file with individual race and ethnicity imputed using the DAS-12.2 and Census 2010 data. To approximate the data used by an expert witness who testified in the court case, we obtain the New York voter file (as of November 16, 2020) from the state Board of Elections. We subset the voters to active voters with addresses in Rockland County, where ERCSD is located. Using the R package `censusxy`, which interfaces with the Census Bureau's batch geocoder, we match each voter to a block and subset the voters to those who live within the geographic bounds of ERCSD [23, 24]. This leaves 58,253 voters, for whom we impute races using the same machinery behind the R package `wru` [25], as described in [19]. This process nearly exactly mimics the one used in the original case.

We examine how the predictions of individual race and ethnicity based on the 2010 Census and DAS-12.2 data result in different redistricting outcomes. Figure 11 compares these two predictions using the proportions of (predicted) Whites, Black, and Hispanic registered voters for each Census block. We find that the predictions based on the DAS-12.2 tend to produce blocks with more White voters than those based on the original



**Figure 11:** Imputed Racial Registrants by Census Blocks. The x-axis represents the percent of a group, as measured by the most likely race from racial imputation using the Census 2010 data. The y-axis represents the corresponding imputation using the DAS-12.2 data.

**Table 2:** East Ramapo MMDs under Census 2010 and DAS-12.2 data. The noise introduced in the DAS-12.2 leads us to undercount the number of majority minority districts in many plans, but never to overcount them.

Census 2010	Number of MMDs from DAS-12.2				Plans
	0	1	2	3	
0	<b>100%</b>	0	0	0	2
1	2	<b>98</b>	0	0	3,581
2	2	40	<b>59</b>	0	6,311
3	6	76	18	<b>0</b>	106

Note: Percentages add to 100% by row.

Census data. As a consequence, the predicted proportions of Black and Hispanic registrants are much smaller, especially in the blocks where they form a majority group.

The precise reason for these biases is unclear. The DAS tends to introduce more error for minority groups than for White voters, and even more error for voters who are in a minority group for their Census block, which is more common for minority voters as well. This additional noise, when carried through a nonlinear transformation such as the Bayes’ rule calculation for racial imputation, may introduce some bias. In addition, the large bias for White and Black voters relative to Hispanic voters suggests that the similarity of surnames between the White and Black populations, compared to the Hispanic population, may also be a factor. Regardless, it is clear that the DAS-injected noise differentially biases voter race imputations at the block level. This pattern may not always yield greater inaccuracies when aggregated to the statewide level—as seen in the prior section—but it is especially prevalent within the ERCSD.

We next investigate whether these systematic differences in racial prediction lead to different redistricting outcomes. Specifically, we simulate 10,000 redistricting plans using DAS-12.2 population and a 5% population parity tolerance. We find that the systematic differences in racial prediction identified above results in the underestimation of the number of MMDs in these plans. As in the original court case, an MMD is defined as a district, in which more than 50% of its registered voters are either Black or Hispanic. Table 2 clearly shows that the number of MMDs based on the DAS-12.2 data never exceeds that based on the 2010 Census for all simulated plans. For example, among 6,311 plans that are estimated to yield 3 MMDs according to the Census data, nearly 60% of them are predicted to have 2 MMDs.

While one should not extrapolate from this single case study, our analysis implies that in small electoral districts such as those of school board elections, the DAS can generate bias that may favor one racial group over another. Although the number of MMDs is underestimated under the DAS data in this case, the

direction and magnitude of racial effects are difficult to predict, as they depend on how the choice of tuning parameters in the DAS algorithm interact with a number of geographical and other factors. At a minimum, this poses a serious challenge in ensuring the effective number of MMDs using DAS-protected data.

## 8 Recommendations

These empirical findings lead to our primary recommendation: release Census P.L. 94-171 data without using the current Disclosure Avoidance System (DAS), and instead rely on a swapping method similar to that applied to the 2010 Census data. Over the past half century, the Supreme Court has firmly established the principle of *One Person, One Vote*, requiring states to minimize the population difference across districts based on the Census data. Our analysis shows that the DAS makes it impossible to follow this basic principle. The only solution is to make Census-block populations invariant, but doing so within the current DAS would, in the Bureau's own admission, require injecting far too much noise into Census tabulations other than total population [26].

We also find that the DAS introduces partisan and racial biases into local data, which may aggregate into large and unpredictable biases at the state level. Since many federal and local elections have narrow margins of victory, relatively small changes to the Census data can result in redistricting plans that produce favorable electoral outcomes for certain candidates and parties. Similarly, these changes affect the number of majority minority districts, either hampering or artificially inflating the voting power of minority groups.

One may argue that the protection of privacy is a worthy cause, and outweighs these potentially negative consequences. Unfortunately, the DAS algorithm fails to universally protect respondent privacy. We are able to predict the individual race of registered voters at least as accurately using the DAS-protected data as when using the original Census data. In sum, we find that the DAS negatively impacts the redistricting process and voting rights of minority groups without providing clear benefits.

If the Census Bureau decides to apply the current DAS to Census P.L. 94-171 data, our recommendation is to increase the privacy loss budget and allocate the increase to improving redistricting outcomes. In addition, the Bureau may consider publishing fewer block-level cross-tabulations in other Census products to ensure more accuracy in the P.L. 94-171 files. In allocating any increased privacy loss budget, we recommend minimizing the change in population at the voting tabulation district (VTD) level. Ensuring that population is accurate at this off-spine geography would help minimize population deviations among the overwhelming majority states which rely on these geographies to draw their districts. This would not fix the problem of ensuring near-exact population equality, but it would help to minimize extreme outliers. In our VTD-level population tabulations, we find that there is around a 1% average deviation in the DAS-12.2 data compared to the 2010 Census data. We recommend aiming for at most a 0.1% average deviation.

Furthermore, we recommend adjusting the parameters of the DAS to address the current demonstrated bias against racially integrated, diverse blocks, and low-turnout areas. Without more detail on the current parameters and workings of the DAS post-processing system, it is difficult to provide more specific recommendations. However, it is vital for the Bureau to ensure that it is not injecting racial and partisan bias into the privacy-protected data.

Finally, should the DAS be used, the Bureau should publish additional information on the known inaccuracies. The current information provided by the Census Bureau with the April PPMF data release provides only marginal distributions of variables, with a focus on total population data. For example, root mean squared error (RMSE) for urban and rural block populations is reported, but these statistics are not cross-tabulated by race or other relevant variables. Reports on inaccuracies and impossibilities must reflect the important relationship that this data has with race, age, population density, and total population. The burden of privacy must not be paid fully by some races or age groups.

## References

- [1] John M. Abowd, Gary L. Benedetto, Simson L. Garfinkel, Scot A. Dahl, Aref N. Dajani, Matthew Graham, Michael B. Hawes, Vishesh Karwa, Daniel Kifer, Hang Kim, Philip Leclerc, Aashwin Machanavajjhala, Jerome P. Reiter, Rolando Rodriguez, Ian M. Schmutte, William N. Sexton, Phyllis E. Singer, and Lars Vilhuber. The modernization of statistical disclosure limitation at the U.S. Census Bureau. 2020.
- [2] Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. Differential privacy and census data: Implications for social and economic research. *AEA papers and proceedings*, 109:403–408,

- 2019.
- [3] David Van Riper, Tracy Kugler, and Steven Ruggles. Disclosure avoidance in the Census Bureau’s 2010 demonstration data product. In *Privacy in Statistical Databases*, Lecture Notes in Computer Science, pages 353–368. Springer International Publishing, Cham, 2020.
  - [4] Jowei Chen and Jonathan Rodden. Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quarterly Journal of Political Science*, 8(3):239–269, 2013.
  - [5] Daniel Carter, Gregory Herschlag, Zach Hunter, and Jonathan Mattingly. A merge-split proposal for reversible Monte Carlo Markov chain sampling of redistricting plans. *arXiv preprint arXiv:1911.01503*, 2019.
  - [6] Daryl DeFord, Moon Duchin, and Justin Solomon. Recombination: A family of Markov chains for redistricting. *arXiv preprint arXiv:1911.05725*, 2019.
  - [7] Eric Autry, Daniel Carter, Gregory Herschlag, Zach Hunter, and Jonathan Mattingly. Multi-scale merge-split Markov chain Monte Carlo for redistricting. *arXiv preprint arXiv:2008.08054*, 2020.
  - [8] Benjamin Fifield, Michael Higgins, Kosuke Imai, and Alexander Tarr. Automated redistricting simulator using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 29(4):715–728, 2020.
  - [9] Cory McCartan and Kosuke Imai. Sequential Monte Carlo for sampling balanced and compact redistricting plans. *arXiv preprint arXiv:2008.06131*, 2020.
  - [10] Christopher T. Kenny, Cory McCartan, Ben Fifield, and Kosuke Imai. *redist: Simulation methods for legislative redistricting*. Available at The Comprehensive R Archive Network (CRAN), 2021.
  - [11] Kevin Fiscella and Allen M. Fremont. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41(4p1):1482–1500, August 2006.
  - [12] Marc N. Elliott, Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. Using the Census Bureau’s surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83, 2009.
  - [13] Kosuke Imai and Kabir Khanna. Improving ecological inference by predicting individual ethnicity from voter registration record. *Political Analysis*, 24(2):263–272, Spring 2016.
  - [14] United States Census Bureau. Meeting redistricting data requirements: Accuracy targets. <https://content.govdelivery.com/accounts/USCENSUS/bulletins/2cb745b>, 2021.
  - [15] Leo A. Goodman. Ecological regressions and behavior of individuals. *American Sociological Review*, 18(6):663–664, 1953.
  - [16] Gary King, Ori Rosen, and Martin Tanner, editors. *Ecological Inference: New Methodological Strategies*. Cambridge University Press, Cambridge, 2004.
  - [17] Orris C Herfindahl. *Concentration in the steel industry*. PhD thesis, Columbia University, 1950.
  - [18] Albert O Hirschman. *National power and the structure of foreign trade*, volume 105. Univ of California Press, 1945.
  - [19] Kosuke Imai and Kabir Khanna. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, pages 263–272, 2016.
  - [20] Ioan Voicu. Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, 5(1):1–13, 2018.
  - [21] United States Census Bureau. Frequently Occurring Surnames from the 2010 Census. [https://www.census.gov/topics/population/genealogy/data/2010\\_surnames.html](https://www.census.gov/topics/population/genealogy/data/2010_surnames.html), 2016.
  - [22] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.
  - [23] Christopher Prener and Branson Fox. *censusxy: Access the U.S. Census Bureau’s Geocoding A.P.I. System*, 2021. R package version 1.0.1.
  - [24] Christopher G. Prener and Branson Fox. Creating open source composite geocoders: Pitfalls and opportunities. *Transactions in GIS*, n/a(n/a), 2021.
  - [25] Kabir Khanna and Kosuke Imai. *wru: Who are You? Bayesian Prediction of Racial Category Using Surname and Geolocation*, 2021. R package version 0.1-12.

- [26] John M. Abowd. Email to James Whitehorne. Exhibit 1 of Plaintiffs' reply in support of combined motion for a preliminary injunction and petition for a writ of mandamus, *Alabama v. U.S. Dep't of Commerce*, No. 3: 21-cv-211-RAH-KFP, 2021.



**16 Vand. J. Ent. & Tech. L. 701**

Vanderbilt Journal of Entertainment and Technology Law  
Summer 2014

**Article**

Jane Bambauer<sup>al</sup> Krishnamurty Muralidhar<sup>aa1</sup> Rathindra Sarathy<sup>aaa1</sup>

Copyright (c) 2014 Vanderbilt Journal of Entertainment & Technology Law, Vanderbilt Law School; Jane Bambauer; Krishnamurty Muralidhar; Rathindra Sarathy

**FOOL'S GOLD: AN ILLUSTRATED CRITIQUE OF DIFFERENTIAL PRIVACY**

**Abstract**

*Differential privacy has taken the privacy community by storm. Computer scientists developed this technique to allow researchers to submit queries to databases without being able to glean sensitive information about the individuals described in the data. Legal scholars champion differential privacy as a practical solution to the competing interests in research and confidentiality, and policymakers are poised to adopt it as the gold standard for data privacy. It would be a disastrous mistake.*

*This Article provides an illustrated guide to the virtues and pitfalls of differential privacy. While the technique is suitable for a narrow set of research uses, the great majority of analyses would produce results that are beyond absurd--average income in the negative millions or correlations well above 1.0, for example.*

*The legal community mistakenly believes that differential privacy can offer the benefits of data research without sacrificing privacy. In fact, differential privacy will usually produce either very wrong research results or very useless privacy protections. Policymakers and data stewards will have to rely on a mix of \*702 approaches--perhaps differential privacy where it is well suited to the task and other disclosure prevention techniques in the great majority of situations where it isn't.*

**Table of Contents**

I. What Is Differential Privacy?	707
A. The Problem	708
B. The Birth of Differential Privacy	712
C. The Qualities of Differential Privacy	717
II. Stunning Failures in Application	720
A. The Average Lithuanian Woman	721
B. Averages of Variables With Long Tails	725
C. Tables	731
D. Correlations	734
III. The Golden Hammer	738
A. Misinformed Exuberance	739
B. Willful Blindness to Context	744
C. Expansive Definitions of Privacy	747
D. Multiple Queries Multiply the Problems	749
E. At the Same Time, Limited Definitions of Privacy	750
F. Difficult Application	752

**Introduction**

A young internist at the largest hospital in a midsized New England city is fretting. She has just diagnosed an emergency room patient with Eastern Equine Encephalitis Virus (EEEV). The diagnosis troubles the internist for a number of reasons. Modern medicine offers neither a vaccine nor an effective treatment.<sup>1</sup> Moreover, the internist remembers that a colleague diagnosed a different patient with EEEV three weeks ago and knows that there was a third case a few weeks before that. The disease is transmitted by mosquitos and is not communicable between humans. However, an influx of cases would suggest that the local mosquito population has changed, putting the city's inhabitants at risk. So, the internist is fretting about whether the three cases that have come through the hospital in the last six weeks merit a phone call to the state and national centers for disease control.

**\*703** To aid her decision, the internist decides to query a state health database to see how many cases of the rare disease have occurred in her city in each of the last eight years. Recently, the state health database proudly adopted differential privacy as a means to ensure confidentiality for each of the patients in the state's database.

Differential privacy is regarded as the gold standard for data privacy.<sup>2</sup> To protect the data subjects' sensitive information, differential privacy systematically adds a random number generated from a special distribution centered at zero to the results of all data queries. The "noise"-- the random value that is added--ensures that no single person's inclusion or exclusion from the database can significantly affect the results of queries. That way, a user of the system cannot infer anything about any particular patient. Because the state health department is also concerned about the utility of the research performed on the database, it has chosen the lowest level of noise recommended by the founders of differential privacy. That is to say, the state has chosen the least privacy-protecting standard in order to preserve as much utility of the dataset as possible. When the internist submits her query, the database produces the following output:<sup>3</sup>

**Query = Count of Patients Diagnosed with EEEV within the City**

Year	N	Year	N
2012	837.3	2007	5,019.3
2011	211.3	2006	868.6
2010	794.6	2005	2,820.6
2009	1,587.8	2004	2,913.9
2008	2,165.5		

What is the internist to make of this data?

**\*704** If the internist is unfamiliar with the theory behind differential privacy, she would be baffled by the responses. She would be especially puzzled by the negative and fractional values since people do not tend to be negative or partial.<sup>4</sup> The internist is likely to conclude the responses are useless, or worse, that the system is seriously flawed.

If the internist happens to be familiar with the theory behind differential privacy, she would know that there is a very good chance--to be precise, a 37% chance--that the system is adding over 1,000 points of noise in one direction or the other. However, even knowing the distribution of noise that is randomly added to each cell, the internist has no hope of interpreting the response. The true values could be almost anything. It could be that the city has consistently diagnosed dozens of patients a year with EEEV, rendering her experience little reason for alarm. Or it could be that the true values are all zero, suggesting that there is reason for concern. The noise so badly dwarfs the true figures that the database query is a pointless exercise.

This hypothetical is a representative example of the chaos that differential privacy would bring to most research database systems. And yet, differential privacy is consistently held up as the best solution to manage the competing interests in privacy and research.<sup>5</sup>

Differential privacy has been rocking the computer science world for over ten years and is fast becoming a crossover hit among privacy scholars and policymakers.<sup>6</sup> Lay descriptions of differential privacy are universally positive. Scientific

American promises that “a mathematical technique called ‘differential privacy’ gives researchers access to vast repositories of personal data while meeting a high standard for privacy protection.”<sup>7</sup> Another journal, *Communications of the ACM*, describes differential privacy in slightly more detailed and equally appealing terms:

Differential privacy, which first emerged in 2006 (though its roots go back to 2001), could provide the tipping point for real change. By introducing random noise and ensuring that a database behaves the same--independent of whether any individual or \*705 small group is included or excluded from the data set, thus making it impossible to tell which data set was used--it's possible to prevent personal data from being compromised or misused.<sup>8</sup>

Legal scholars have also trumpeted the promise of differential privacy. Felix Wu recommends differential privacy for some scientific research contexts because the query results are “unreliable with respect to any one individual” while still making it sufficiently reliable for aggregate purposes.<sup>9</sup> Paul Ohm explains differential privacy as a process that takes the true answer to a query and “introduces a carefully calculated amount of random noise to the answer, ensuring mathematically that even the most sophisticated reidentifier will not be able to use the answer to unearth information about the people in the database.”<sup>10</sup> And Andrew Chin and Anne Klinefelter recommend differential privacy as a best practice or, in some cases, a legal mandate to avoid the reidentification risks associated with the release of microdata.<sup>11</sup>

Policymakers have listened. Ed Felten, the chief technologist for the Federal Trade Commission, praises differential privacy as “a workable, formal definition of privacy-preserving data access.”<sup>12</sup> The developers of differential privacy have even recommended using the technique to create privacy “currency,” so that a person can understand and control the extent to which their personal information is exposed.<sup>13</sup>

These popular impressions give differential privacy an infectious allure. Who wouldn't want to maximize database utility while ensuring privacy? The truth, of course, is that there is no simple solution to the eternal contest between data privacy and data utility. As we will show, differential privacy in its pure form is a useful tool in certain \*706 narrow circumstances. Unfortunately, most research occurs outside of those circumstances, rendering a pure form of differential privacy useless for most research. To make differential privacy practical for the vast majority of data research, one would have to diverge significantly from differential privacy's pure form.

Not surprisingly, this is the direction in which advocates of differential privacy have gone.<sup>14</sup> It is the only way to go if one harbors hopes for general application of the technique. But the only way to convert differential privacy into a useful tool is to accept and adopt a range of compromises that surrender the claim of absolute “ensured” privacy. In other words, a useful version of differential privacy is not differential privacy at all. It is a set of noise-adding practices indistinguishable in spirit from other disclosure prevention techniques that existed well before differential privacy burst onto the scene. Thus, differential privacy is either not practicable or not novel. This Article provides a comprehensive, but digestible, description of differential privacy and a study and critique of its application. Part I explains the age-old tension between data confidentiality and utility and shows how differential privacy strives to thread the needle with an elegant solution. To this end, Part I recounts a brief history of the development of differential privacy and presents a successful application of differential privacy that demonstrates its promise. Part II explores the many contexts in which differential privacy cannot provide meaningful protection for privacy without sabotaging the utility of the data. Some of the examples in this section are lifted directly from the differential privacy literature, suggesting, at least in some cases, that the proponents of differential privacy do not themselves fully understand the theory. The most striking failures of differential privacy (correlations greater than 1, average incomes in the negative millions) track some of the most general, common uses of data. Part II demonstrates clearly that differential privacy cannot serve as the lodestar for the future of data privacy. Part III conducts a postmortem. What went wrong in the applications of differential privacy described in Part II? Looking forward, how can we know in advance whether differential privacy is a viable tool for a particular research problem? The answers provide insight into the limitations of differential privacy's theoretical underpinnings.

These limitations can point researchers in the right direction, allowing them to understand when and why a deviation \*707 from the strict requirements of differential privacy is warranted and necessary. We also identify and correct some misinformed legal scholarship and media discussion that give unjustified praise to differential privacy as a panacea.

The Article concludes with a dilemma. On one hand, we praise some recent efforts to take what is good about differential privacy and modify what is unworkable until a more nuanced and messy--but ultimately more useful--system of privacy practices are produced. On the other hand, after we deviate in important respects from the edicts of differential privacy, we end up with the same disclosure risk principles that the founders of differential privacy had insisted needed to be scrapped. In the end, differential privacy is a revolution that brought us more or less where we started.

## I. What Is Differential Privacy?

Protecting privacy in a research database is tricky business. Disclosure risk experts want to preserve many of the relationships among the data and make them accessible.<sup>15</sup> This is a necessary condition if we expect researchers to glean new insights. However, the experts also want to thwart certain types of data revelations so that a researcher who goes rogue-- or who was never really a researcher to begin with--will not be able to learn new details about the individuals described in the dataset. How to preserve the "good" revelations while discarding the "bad" ones is a puzzle that has consumed the attention of statisticians and computer scientists for decades.<sup>16</sup>

When research data sets are made broadly available for research purposes, they usually take one of two forms.<sup>17</sup> Sometimes \*708 the disclosure risk expert prepares and releases microdata--individual-level datasets that researchers can download and analyze on their own. Other times, the expert prepares an interactive database that is searchable by the public. An outside researcher would submit a query or analysis request through a user interface that submits the query to the raw data. The interface returns the result to the outside researcher (sometimes after applying a privacy algorithm of some sort). The techniques for preserving privacy with these alternative research systems are quite different, not surprisingly. The debate over how best to prepare microdata is lively and rich.<sup>18</sup>

The public conversation about interactive databases, in contrast, is underdeveloped.<sup>19</sup> Outside of the technical field, hopeful faith in differential privacy dominates the discussion of query-based privacy.<sup>20</sup> This Part first explains the problem differential privacy seeks to solve. It is not immediately obvious why a query-based research system needs any protection for privacy in the first place, since outside researchers do not have direct access to the raw data; but even an interactive database can be exploited to expose a person's private information. Next, we demystify differential privacy --the creative solution developed by Microsoft researcher Cynthia Dwork --by working through a successful example of differential privacy in action.

### A. The Problem

Six years ago, during a Eurostat work session on statistical data confidentiality in Manchester, England, Cynthia Dwork, an energetic and highly respected researcher at Microsoft, made a startling statement.<sup>21</sup> In a presentation to the world's statistical \*709 privacy researchers, Dwork announced that most, if not all, of the data privacy protection mechanisms currently in use were vulnerable to "blatant non-privacy."<sup>22</sup>

What Dwork meant by "blatant non-privacy" comes from a 2003 computer science publication by Irit Dinur and Kobbi Nissim.<sup>23</sup> Dinur and Nissim showed that an adversary--that is, a malicious false researcher who wishes to expose as much personal information as possible by querying a database--could reconstruct a binary database (a database containing only responses consisting of "0" s and "1" s) if they had limitless opportunity to query the original database, even if noise of magnitude  $\pm$  is

added to the results of the queries, as long as  $E$  is not too large.<sup>24</sup> Dinur and Nissim defined “non-privacy” as a condition in which an adversary can accurately expose 99% of the original database through queries.<sup>25</sup>

To understand how such an attack works, suppose a database contains the HIV status of 400 patients at a particular clinic. The adversary knows that  $E = 2$ , meaning that the noise added or subtracted is no greater than 2. The adversary knows that for any response he receives from the system, the true value is within  $\pm 2$  of the response. Now assume that the adversary issues the query, “How many of the first 20 individuals in the database are HIV positive?” For the sake of argument, let us assume that the true answer to this query is 5. And assume that the system adds 2 to the true answer and responds with 3. Now the adversary asks: “How many of the first 21 individuals in the database are HIV positive?” Assume that the twenty-first individual is HIV positive, and the true answer to this query is 6. The system adds +2 to the true answer and responds with 8. From the response to the first query, the adversary knows that the true answer could not possibly be greater than 5. From the response to the second query, the adversary knows that the true answer could not possibly be less than 6. So, he can correctly conclude that: (a) the twenty-first individual must be HIV positive, and (b) there are 5 HIV positive cases among the first 20 individuals.

There are  $2^{400}$  possible queries of this sort, and if an adversary used all of them, he could correctly reconstruct 99% of the HIV statuses. Dinur and Nissim also showed that even under more realistic scenarios where the number of queries is bounded, and even when the noise added occasionally exceeds  $E$ , an adversary can still recreate a rather accurate database as long as  $E$  is not too large and the value of  $E$  is known.<sup>26</sup>

These results provide important theoretical foundations for disclosure risk because they show that moving from a microdata release to a query system does not automatically assure privacy. A query system must be designed in a thoughtful way. However, from a practical perspective, the consequences of the Dinur-Nissim discovery are not as serious as they seem at first glance. For instance, if the selection of the noise function,  $E$ , is large enough, it can thwart an adversary's attempt to construct a nearly accurate database no matter how many queries he submits.<sup>27</sup>

But the most helpful limitation is the natural bound on the number of queries that a researcher can submit. Even for small databases, like the HIV database described above, an adversary would not be able to issue all of the queries necessary to attempt a full database reconstruction because of the sheer number of queries required. A database with 400 subjects would require  $2^{400}$  queries. To give a sense of scale,  $2^{332.2}$  is a googol, which is greater than the number of atoms in the observable universe.<sup>28</sup>

In addition to these natural limitations of the adversary, a query system may limit the total number of queries issued to the database or impose other restrictions when responding to queries.<sup>29</sup> The data producer can also withhold information about the amount of noise added. Once an adversary is constrained in the number of query submissions, an appropriate selection of noise can virtually guarantee that a reconstruction attack will not work.<sup>30</sup> The Dinur-Nissim attack would also fail if the administrator were to change the values in the original database and use the modified database to respond to all queries.<sup>31</sup>

Reconstruction attacks are not the only privacy threats that concern data providers. If an adversary can accurately figure out one highly sensitive attribute of a single data subject, such as an HIV diagnosis, the revelation would be disconcerting, even if the rest of the original database remained unknown. Meanwhile, data providers might shrug at a 99% accurate candidate database constructed by an “adversary” who guessed that everybody in the database had a negative HIV status.<sup>32</sup>

Thus, disclosure risk experts have long understood that the best approach to protecting privacy is one that is contextually sensitive.<sup>33</sup> Privacy risks fall disproportionately on data subjects whose demographics or other characteristics make them unusual.<sup>34</sup> Disclosure risk experts traditionally employ a range of techniques to protect outlier data subjects and highly sensitive attributes. Most of the time, for the sake of simplicity and ease of application, a database query system will add some random noise to the results generated by a particular query, and that noise usually falls within some bounded range.<sup>35</sup> That way, the

utility of the response is not swamped by the noise added at the end. The disclosure limitation community was <sup>36</sup> interested in developing alternatives to these common noise-adding practices when Dwork made her provocative presentation.

The holistic approach was unsatisfying to Dwork. She criticized the popular approaches for being “syntactic” and context driven.<sup>37</sup> Instead, Dwork insisted that the practical compromises were not necessary. One could design a query system that avoids even the theoretical risks of query attacks, or, rather, allows the theoretical risks only within a predefined range of tolerances.

## B. The Birth of Differential Privacy

Differential privacy does two important things at once. First, it defines a measure of privacy, or rather, a measure of disclosure--the opposite of privacy.<sup>38</sup> And second, it allows data producers to set the bounds of how much disclosure they will allow.<sup>39</sup> For Dwork, if, based on a query result--or a series of results--an adversary can improve his prediction of a person's attributes, then any such improvement in the prediction represents a disclosure.<sup>40</sup>

In its purest form, this definition is too strong to be usable in settings where disclosure is strictly prohibited.<sup>41</sup> It obliterates research utility. Suppose, for example, an adversary has external knowledge that a particular person, Claire, is female. Now, any research describing gender differences along various dimensions would improve his predictions of Claire's attributes. While his best guess at her income would have been the average US income in the absence of better information, his prediction would be improved (though still not good) by learning that women earn less, on average, than men do. If disclosure were defined this broadly, every published statistic would violate privacy.

Dwork avoided this absurdity by proposing an elegant solution: differential privacy ensures that the presence or absence of an <sup>42</sup> individual does not significantly affect the responses that the system provides. More precisely, differential privacy disclosure occurs when, for any individual, the probability that a query will return a particular result in the presence of that individual in the database differs from the probability that a query would return that same result in the absence of that individual.<sup>43</sup> The measure of the disclosure for a particular query to a particular individual is the ratio of those two probabilities--the probabilities that the query system would return the result with, and then without, the individual's data.<sup>44</sup> Ideally, this ratio would be one, allowing no disclosure at all. But since this is impossible to achieve if the responses are to be useful, the data curator can select some small level of disclosure that society is willing to tolerate. The closer to one the ratio is, the less disclosure has taken place.<sup>45</sup>

For a query system to satisfy differential privacy, the system must add noise that ensures it only returns results such that the disclosure for everybody stays within certain predetermined bounds.<sup>46</sup>

Consider this example: Suppose a data producer had made differential privacy commitments, promising that the ratio of probabilities for all possible people and all possible values of return results would never be less than 1/2 or more than 2. And suppose that the database contains the wealth for the year 2010 for all Americans whose primary residence is in the state of Washington. An adversary submits the query, “How many people have more than \$1 million in wealth?”

Suppose the true answer is 226,412, and one of those millionaires is Bill Gates.<sup>47</sup> The query system will apply some noise randomly drawn from a distribution, but what should that distribution be? Well, it must be drawn such that it does not diverge too greatly from the distribution of responses if the database didn't include Bill Gates. Removing Bill Gates from the database, the answer to the query is 226,411, and noise from the same distribution is randomly drawn to apply to that number instead. The query system must use a distribution that ensures that when we look at the probability of all possible returned results based

on the true result or \*714 the result with a record deleted, the distributions are not too far apart. Figure 1 plots the distribution that has this quality.

### Figure 1--Distribution of Query Response if the True Answer Contains, or Does Not Contain, Bill Gates

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

Reflect for a moment on the reasons that we want the query system to produce similar results whether Bill Gates is or is not in the query system. Most people know perfectly well that Bill Gates lives in Seattle and is a billionaire, so they would not be surprised to discover that he is included in the count of millionaires. But suppose an eccentric adversary knew the identity of every millionaire in Washington except Bill Gates. Suppose also that he knew that everybody except the 226,411 millionaires and Bill Gates were not millionaires. The only thing he does not know is whether Bill Gates has at least \$1 million. If this adversary is clever, and if the data producer had used bounded noise, the adversary might be able to improve his inference that the noise centers around 226,411 (suggesting Gates is not a millionaire) or around 226,412 (suggesting that he is a millionaire).<sup>48</sup> Differential privacy ensures that the \*715 system does not produce answers that behave very differently under either case.

Mathematically, the promise of differential privacy looks like this:

Given a database  $X$ , and a hypothetical database  $X^*$  that differs from  $X$  by the deletion or addition of just one record, differential privacy ensures that<sup>49</sup> <<equation>>

The data producer gets to choose <<unknown symbol>>, and the choice of << unknown symbol>> will determine how much disclosure (as defined by Dwork and described above) the system will tolerate. The reason for the use of  $e$  (2.71828 . . .) is that by setting up the differential privacy promise this way, it corresponds precisely with a distribution curve already well known to statisticians--the Laplace distribution curve.<sup>50</sup> Laplace distribution has precisely the quality we are looking for: when the curve is shifted over a certain amount, the ratio of probabilities for the original and shifted curve stay within a predesignated boundary.

To employ differential privacy, a data curator would do the following:

- (1) Select <<unknown symbol>>. The smaller the value, the greater the privacy.
- (2) Compute the response to the query using the original data. Let  $a$  represent the true answer to the query.
- (3) Compute the global sensitivity ( $\Delta_{\pm}$ ) for the query. Global sensitivity is determined by answering the following: "Assume that there are two databases  $X$  and  $X^*$  which differ in exactly one record and that the answer to this query from database  $X$  is  $a$  and that from database  $X^*$  is  $a^*$ . For any two such databases  $X$  and  $X^*$  in the universe of all possible databases for the queried variable, what is the maximum possible absolute difference between  $a$  and  $a^*$ ?"<sup>51</sup> According to \*716 to Dwork and Smith, "The sensitivity essentially captures how great a difference (between the value of  $\pm$  on two databases differing in a single element) must be hidden by the additive noise generated by the curator."<sup>52</sup> If the noise can protect this difference, then of course, all other, smaller, differences will also be protected. This is the key to differential privacy's protection.
- (4) Generate a random value (noise) from a Laplace distribution with mean = 0 and scale parameter <<unknown symbol>>. Let  $y$  represent the randomly generated noise.
- (5) Provide the user with response  $R = a + y$ . The noise added ( $y$ ) is unrelated to the characteristics of the actual query (number of observations in the database or query and the value of the true response) and is determined exclusively by  $\Delta_{\pm}$  and <<unknown symbol>>.<sup>53</sup>

Observe this as applied to the example of the number of millionaires in Washington. The data producer wanted the ratio of responses to stay within  $1/2$  and  $2$  when a person's information was included or removed from the database. Therefore, the data producer selected  $\epsilon = \ln(2)$ .<sup>54</sup> The global sensitivity here has to be one. Since the query asks for a headcount, the greatest difference any single person can make to the count is one.

We know that the true answer to the query is 226,412. We do not know what answer the data query system will produce because it takes the true answer and adds some randomly chosen noise from a Laplace distribution. But we can look at the range of responses such a system produces. Figure 2 plots the chance of seeing any particular response.

**\*717 Figure 2--Number of Millionaires in Washington State  $\epsilon = \ln(2)$   $\Delta f=1$**

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

As you can see, differential privacy works quite well here. The query system produces results that tend to provide utility--the responses are very unlikely to be too far off from the true answer--and the system also insures against disclosure. This is a true win-win.

### C. The Qualities of Differential Privacy

Much of this Article is devoted to illuminating the defects of differential privacy, but we do not want the reader to walk away without an understanding of its virtues. As the millionaires example demonstrates, Dwork's measure of disclosure makes the issue of auxiliary information easy to handle and potentially very privacy protecting. Even if the adversary knows everything in the database except one particular piece of information, differential privacy assures that the responses from the database--in the presence or absence of this record--are indistinguishable within a factor of  $\epsilon$ . If we have confidence that this factor is small enough to be considered safe, then we need not speculate about what a user's motives are or how much information he already has. He can be a super-adversary, knowing almost everything, and his efforts will still be frustrated.

**\*718** Differential privacy also protects against possible inferences based on a person's absence from a database.<sup>55</sup> A person's absence might reveal something very important. To see why this is so, return to the example of the income data for Washington residents. This time let us assume that the adversary's target is Larry Page, who does not live in Washington--and thus would not be in the database. If the last piece of information that the adversary needed about Larry Page was whether or not he lived in Washington, and the adversary also knew all of the 226,412 millionaires in Washington, then the fact that noise is not centered around 226,413 would reveal to the adversary that Larry Page does not live in Washington, and a disclosure would occur.

Dwork consciously made some overt choices and sacrifices when she developed differential privacy. For one thing, as Dwork herself has noted, microdata releases cannot be prepared in a way that strictly complies with differential privacy, so the standard applies only to query systems.<sup>56</sup> Also, much rides on the query designer's selection of  $\epsilon$ . The smaller it is, the more privacy protecting, but also the more utility damaging since the noise added will tend to be larger.<sup>57</sup> Therefore, we must rely on the judgment of the data producer to select an appropriate  $\epsilon$  that strikes the right bargain between privacy and utility.<sup>58</sup> This selection is all the more difficult because, whatever selection the data producer chooses for the system's overall privacy protections ( $\epsilon$ ), he must also decide how many queries researchers are allowed to make. Because the effects of successive queries on disclosure are cumulative, the data producer will have to divide his choice of  $\epsilon$  by the anticipated number of queries.<sup>59</sup>

**\*719** Finally, in defining disclosure as she does, Dwork implicitly rejects other definitions of disclosure that would disclose families or groups.<sup>60</sup> Dwork ensures that an individual is not distinguishable from the results of a query, but she does not build in protections against revelations for families or subgroups.<sup>61</sup> What differential privacy can promise is that "the ability of an adversary to inflict harm (or good, for that matter)--of any sort, to any set of people--should be essentially the same, independent



of whether any individual opts in to, or opts out of, the dataset.”<sup>62</sup> For most research applications, this distinction between individuals and groups make sense.<sup>63</sup> After all, a research study finding that smoking causes cancer says something about every person who smokes--it allows an adversary to predict with better accuracy whether a particular smoker (whether they were in the research database or not) has cancer. But the adjustment to the adversary's prediction about that particular smoker would be based on group phenomena and not on individualized information about this particular smoker.<sup>64</sup>

Nevertheless, some data producers may be concerned about family and group disclosures. Some group disclosures--like whether a family has a congenital disease--might be more important than protecting against the theoretical possibility that somebody might not \*720 know that Bill Gates lives in Washington. If so, they will have to rely on techniques beyond differential privacy.

## II. Stunning Failures in Application

All database query systems serve the purpose of providing reasonably accurate information. Research results are the *raison d'être* for the query system in the first place. Inaccurate responses can be useless. In some cases, they can be positively harmful. Privacy is trivially easy to achieve if the data producer has no minimum standards for response accuracy. Responding to all queries with “0” would do the trick. Yet to facilitate useful research, maintaining reasonable accuracy has to be a priority. Unfortunately, differential privacy has great difficulty performing under most realistic conditions. The illustrations in this Part show that a data producer who wishes to comply with differential privacy will almost always have to choose between adding so much Laplace noise that the query results are ludicrous or adding so little noise that the dataset is left vulnerable to attack. DPA1#There are exceptions--the Washington millionaires example from the previous part is one of them. In Part III, this Article will explain when differential privacy can work. But first, let us examine how differential privacy can quickly go off the rails. As in most illustrations of differential privacy, we assume that the curator or administrator of the database allows for only one query to the database. This assumption is completely unrealistic since thousands (or perhaps millions) of queries may be issued to the database.<sup>65</sup> When the database receives many queries, the privacy afforded is diminished by each individual query.<sup>66</sup> We will consider this issue in more detail in Part III. The assumption of a single query presents differential privacy in the best possible light. Considering multiple queries means that the noise added will increase as a direct multiple of the number of queries, making matters much worse.<sup>67</sup>

### \*721 A. The Average Lithuanian Woman

One of the most frequently cited examples to justify the need for differential privacy is also, in our view, one of the most misguided. Dwork presents this example as she contemplates the disclosure risk from a database that includes the heights of Lithuanian women:

Finally, suppose that one's true height is considered sensitive. Given the auxiliary information “[Alan] Turing is two inches taller than the average Lithuanian woman,” access to the statistical database teaches Turing's height. In contrast, anyone without access to the database, knowing only the auxiliary information, learns much less about Turing's height.<sup>68</sup>

The idea is that even individuals who are not represented in the database stand to suffer a privacy violation.<sup>69</sup> Therefore, to set up the problem, we assume that (1) Alan Turing's height is not known to the public; (2) the height of the average Lithuanian woman is available only to those who have access to the query database; and (3) the auxiliary information that Turing is two inches taller than the average Lithuanian woman is known to the adversary.

This is an odd hypothetical. After all, in order to create the auxiliary information that “Turing is two inches taller than the average Lithuanian woman,” the creator of the information must know both Turing's height and the height of the average Lithuanian woman. This would have to be Turing himself or somebody privy to his sensitive height information; but then, how did they know the height of Lithuanian women? Even if a data curator is determined to protect height information, this particular style of auxiliary information falls outside the set of risks that differential privacy is designed to reduce.<sup>70</sup> The meat of the sensitive information is contained in the auxiliary information. The auxiliary information is the disclosure--it is just communicated in reference to some external fact.<sup>71</sup>

In any case, let us humor the hypothetical. What would differential privacy tell the curator of a database about the height of Lithuanian women to do in order to protect the privacy of Alan Turing--and others? Let us follow the steps laid out in Part I.

**\*722** 1. Select <<unknown symbol>>

First, the curator of the database containing the height of Lithuanian women must decide on the value of <<unknown symbol>> (the acceptable level of disclosure). The curator must make a judgment call on how far off the probability distributions are allowed to be when the database does, and does not, include a particular person. Dwork has suggested that <<unknown symbol>> is often in the order of 0.01 or 0.1, “or in some cases,  $\ln 2$  or  $\ln 3$ .”<sup>72</sup> Since the primary objective in this exercise is to prevent disclosure, we should use a fairly high privacy standard, setting <<unknown symbol>> = 0.1. (Remember, the smaller the <<unknown symbol>>, the greater the noise).

The query “What is the height of the average Lithuanian woman” is actually two queries rolled into one because it requires two different pieces of information: the number of Lithuanian women and their total height. Further, since <<unknown symbol>> = 0.1 and the response involves two different queries, for each query, we will set <<unknown symbol>> = 0.05.

2. Compute the Response to the Query Using the Original Data

According to Statistics Lithuania, the population of Lithuania in 2012 was just over 3 million, with females accounting for approximately 1.6 million.<sup>73</sup> The average height of Lithuanian women is 66 inches.<sup>74</sup>

3. Compute the Global Sensitivity ( $\Delta f$ ) for the Query

We must determine global sensitivity for both the count of Lithuanian women and the sum of their heights. The absence or presence of an individual will change the number of Lithuanian woman by exactly one and hence  $\Delta f = 1$ . But how about the sum of the height query? The largest difference in the sum of heights between any two databases that differ in one record would occur when one database contains the tallest living person and the other does not. The difference in the total height between the two databases would equal the height of the tallest living person. The height of the tallest person living in the world today is 99 inches (8'3”), so  $\Delta f$  for the sum of the height query is 99.

**\*723** 4. Generate a Random Value (Noise) from a Laplace Distribution with Mean = 0 and Scale Parameter <<equation>>

Based on the information worked out above, the Table provides the original answers, the noise added, and the response to a query operating on the entire population of Lithuanian women.

**Table 1--Response to Query on Average Height Over Database of Lithuanian Women <<unknown symbol>> = 0.05**

	True values	$\Delta f$	Laplace Noise		Noise Added Response	
			Low (0.01)	High (0.99)	Low	High
# of Lithuanian Women	1,603,014	1	78	78	1,602,936	1,603,092
Total Height (inches)	105,798,924	99	7,746	7,746	105,791,178	105,806,670

Average Height (inches)	66	65.99	66.01
-------------------------	----	-------	-------

Because this query analyzes over one million people, the large n keeps the Laplace noise from drowning out the true signal. Thus, the low estimate of average height is within 0.02” of the high estimate for average height. Anyone who knows that Turing is 2” taller than the average Lithuanian woman will have no trouble concluding that he is 68” tall, even after the data curator adopts the precautions of differential privacy.

However, the decision to adopt differential privacy to protect everyone (including Turing and the world's tallest person), whether or not they are in the database, comes at a very high cost in other contexts. What if the adversary knew that Turing was 2” taller than the average woman in the small Lithuanian town of Smalininkai (population 621, of whom 350 are women)? Or what if the adversary knows Turing is 2” taller than the average employed woman in Smalininkai? Now, to protect the possibility of disclosure for Turing (as well as the world's tallest person), the query system must allow the possibility of inventing a land of 30-foot-tall women. It also may produce tiny towns with people measuring less than 1” tall. Tables 2 and 3 display the range of results for average heights of these smaller subpopulations, using the same differential privacy parameters we set before.

**\*724 Table 2--Response to Query on Average Height of Smalininkai Women Over Database of Lithuanian Women <<unknown symbol>> = 0.05**

	True values	$\Delta f$	Laplace Noise		Noise Added Response	
			Low (0.01)	High (0.99)	Low	High
# of Smalininkai Women	350	1	78	78	272	428
Total Height (inches)	23,100	99	7,746	7,746	15,354	30,846
Average Height (inches)	66				35.9	113.5

**Table 3--Response to Query on Average Height of Employed Smalininkai Women Over Database of Lithuanian Women <<unknown symbol>> = 0.05**

	True values	$\Delta f$	Laplace Noise		Noise Added Response	
			Low (0.01)	High (0.99)	Low	High
# of Employed Smalininkai Women	120	1	78	78	42	198
Total Height (inches)	7,920	99	7,746	7,746	174	15,666
Average Height (inches)	66				0.88	375.1

Notice that the distributions of noise that the equation adds to the count and total heights in Tables 2 and 3 are identical to the distributions shown in Table 1. This should not be surprising, since the shape of the noise distribution is determined solely by the values of <<unknown symbol>> and  $\Delta f$ . These values did not change since we still have to protect the world's tallest person. However, while the noise was relatively small as applied to the entire female population of Lithuania, the same noise quickly overwhelms the true values when taking the averages over smaller subpopulations.

One could rationalize that smaller subgroups need more noise to protect the confidential information. However, research databases often rely on randomly selected subsamples of the population to avoid the significant costs of surveying every person. The database applies the exact same distribution of noise to an unknown, random \*725 subsample of the population. So, if a world census allowed researchers to query average heights on a randomly selected sample of 120 Lithuanian women, the results would look just as bizarre as the ones reported in Table 3.

Matters would be much worse if we assume that the curator decides to respond to several hundred or thousands of queries. The noise currently added is large enough to overwhelm the true answer; with one thousand queries, the noise added to comply with differential privacy standards would increase a thousand fold!<sup>75</sup>

**B. Averages of Variables With Long Tails**

Differential privacy has the potential to radically distort averages of variables (like height) that are normally distributed, but the distortion is even worse on variables like income that have a skew--that is, where some members of the population have values that are very distant from the median. For instance, while the median family income in the United States is just under \$53,000,<sup>76</sup> a few hedge fund operators like George Soros have income exceeding \$1 billion.<sup>77</sup> Scholars often refer to these distant values to as the “long tail” of the distribution.

Booneville, Kentucky, is a small and struggling town.<sup>78</sup> Its population is just over 100, and the median household income is just above the poverty line.<sup>79</sup> Suppose the town decided to make a database available for public research as part of a new transparency initiative designed to inspire research on public welfare and the prevention of poverty. Under normal circumstances, one might counsel the town to include only a random subsample of residents and to join forces with other similar towns so that a data user might not be able to discern the precise town in which the data subjects live. There may be other precautions too, based on the context and nature of the data. But in this hypothetical scenario, the town has opted instead to rely on differential privacy. After all, one of the core strengths of **\*726** differential privacy is that the methods of masking query responses are completely independent of the size and nature of the Booneville data--the town can have mathematical certainty of meeting privacy standards regardless of the particular features of its town.<sup>80</sup>

What happens when a researcher queries the average income of Booneville residents? In this case, income is the confidential variable; we do not want an adversary to be able to tell something about his target--either about his income or using his income--based on what he learns from the response to the query. In particular, the town would need to ensure that the adversary would not be able to rule out that his target--a Booneville resident--is a billionaire. After all, when large values are included in an analysis of the mean, the outlier has an outsized effect on the analysis. So a reported mean that roughly matches the incomes of the rest of the Booneville population would suggest that the last person in the sample is not a billionaire. Also, the town might need to ensure that an adversary who knows everything about George Soros except where he lives is not able to rule out Booneville as George Soros's hometown. Thus, even if the highest income among Booneville residents is \$50,000, the probability of any particular response coming back from the query needs to be not so far off from the probability that that response would come back if George Soros lived in Booneville.<sup>81</sup> That is the promise of differential privacy. Unfortunately, this privacy promise also means that the response is likely to be useless.

Now, we will work through the application following the instructions we provided in Part I.

### 1. Select <<unknown symbol>>

First, the town must decide how much disclosure it is willing to tolerate and will have to allocate this disclosure among all the queries it issues to this database. For simplicity we will assume that the town will use <<unknown symbol>> = 0.50 for this particular query.<sup>82</sup>

### **\*727** 2. Compute the Response to the Query Using the Original Data

Suppose, for this illustration, the true per capita income for Booneville residents is \$23,426 (which is the value reported by the US Census Bureau's FactFinder web tool for 2007-11).<sup>83</sup>

### 3. Compute the Global Sensitivity ( $\Delta f$ ) for the Query

As we saw with the example of Lithuanian women, this query actually involves two separate global sensitivities (sum of income and count of people), but we will take a shortcut by dividing the global sensitivity for income by the number of data subjects responsive to the query.<sup>84</sup> In this case, only 59 Booneville residents were in the workforce according to FactFinder.<sup>85</sup>

When it comes to income, the global sensitivity is very large. It is the difference between the highest-paid man in the world and an unemployed man. For the sake of illustration, we will assume that the highest income is \$1 billion and the lowest is \$0. Thus, the global sensitivity is \$1 billion.<sup>86</sup>

4. Generate a Random Value (Noise) from a Laplace Distribution with Mean = 0 and Scale Parameter  $\epsilon$

Now comes the fun part--the selection of noise to add to the true answer (\$23,426). A Laplace distribution randomly selects noise, but the reason we went through all the work of determining the global sensitivity and the value of  $\epsilon$  is that these two factors determine the distribution--the likelihood of how much noise the equation adds. To satisfy differential privacy, the Laplace distribution which randomly selects the noise must have a standard deviation of  $\epsilon$  million.

Thus, although the true answer to the query "What is the average income of the inhabitants of Booneville?" is \$23,426, the answer after the differential privacy process is very likely to be over \$10 million.<sup>87</sup> It is also very likely to come out lower than negative \$10 million. In fact, the chance that the query answer will be within \$1 million of the true answer is under 3%.<sup>88</sup>

Table 4 and Figure 3 show the Laplace distribution of noise. The two dotted lines represent negative \$5 million and \$5 million. The small area between the dotted lines visually represents the chance that the noise would fall within that range.

**Table 4--Distribution of Noise Added to a Query for Average Income Where the True Answer is \$23,426  $\epsilon = 0.5$ ,  $\Delta f = \$1$  Billion**

Noise Level	Noise Added	Response (True Value + Noise)
Very Low (0.001)	210,664,681	\$210,641,255
First percentile (0.01)	132,610,949	\$132,587,523
Fifth percentile (0.05)	78,053,732	\$78,030,306
Tenth percentile (0.10)	54,557,217	\$54,533,791
Twenty-fifth (0.25)	23,496,515	\$23,473,089
Fiftieth (0.50)	0	\$23,426
Seventy-fifth (0.75)	23,496,515	\$23,519,941
Ninetieth (0.90)	54,557,217	\$54,580,643
Ninety-fifth (0.95)	78,053,732	\$78,077,158
Ninety-ninth (0.99)	132,610,949	\$132,634,375
Very High (0.999)	210,664,681	\$210,688,107

**\*729 Figure 3--Distribution of Noise Added to a Query for Average Income  $\epsilon = 0.5$ ,  $\Delta f = \$1$  Billion**

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

Table 5 shows the distribution of noise under various choices of  $\epsilon$ . Even if the data producer chose 1 for the value of  $\epsilon$ , a choice that might garner criticism for being insufficiently protective of privacy, the response to any query on the income variable would be swamped by noise.

**Table 5--The Probability that Laplace Noise Will Be Selected from Specified Ranges  $\Delta f = \$1$  Billion**

	$\epsilon = 0.01$	$\epsilon = 0.10$	$\epsilon = 0.50$	$\epsilon = 1.00$	$\epsilon = \ln(3)$
$\pm 10,000$	0.0000	0.0001	0.0003	0.0006	0.0006
$\pm 100,000$	0.0001	0.0006	0.0029	0.0059	0.0065
$\pm 500,000$	0.0003	0.0029	0.0146	0.0291	0.0319
$\pm 1$ Million	0.0006	0.0059	0.0291	0.0573	0.0628
$\pm 5$ Million	0.0029	0.0291	0.1371	0.2555	0.2768
$\pm 10$ Million	0.0059	0.0573	0.2555	0.4457	0.4770

±100 Million	0.2555	0.9477	1.0000	1.0000	1.0000
±1 Billion	0.4457	0.9973	1.0000	1.0000	1.0000

Table 5 also reveals another important fact about differential privacy method; by design, the noise added to a query is entirely independent from the values of the database. The Laplace noise <sup>89</sup> distribution is determined by global sensitivity and the choice of  $\epsilon$ , neither of which required the data producer to consult the database. <sup>89</sup> The noise is independent from the actual answer to the query. <sup>90</sup> So Table 5 represents the noise that would be added not only to this hypothetical query involving a small town in Kentucky but to any analysis of income over data this size. Therefore, if the US Census Bureau chose to adopt differential privacy in an online query system for the Current Population Survey, it too would add and subtract hundreds of millions in noise to protect George Soros when a user queried, “What is the average income for employed females over the age of 65 living in the South Bronx?” Note that this applies even to queries about females because the last pieces of information an adversary might need about George Soros is that he is not an older female living in the Bronx.

When it comes to the analysis of continuous, skewed variables (like income), differential privacy's strict and inflexible promises force a data producer to select from two choices: he can either obliterate the data's utility or he can give up on the type of privacy that differential privacy promises.

For comparison's sake, let us look at how the Census Bureau's American FactFinder service actually reports the income of the residents in Booneville, Kentucky. <sup>91</sup> According to American FactFinder, the average income of the 51 working individuals in Booneville is \$21,907 and a margin of error of  $\pm$  \$11,247. <sup>92</sup> For any realistic selection of  $\epsilon$ , this release of information by the Census Bureau would violate differential privacy since an adversary would be able to conclude that it is extremely unlikely that anyone living in Booneville has an income of \$1 billion. From the first line of Table 5 above, one can see that the probability of observing a differentially private response within the range that the Census Bureau has released is infinitesimally small.

It is hard to fault the Census Bureau for not using differential privacy. After all, a little external information and knowledge of the world would suggest that it is extremely unlikely that a multi-billionaire lives in a small, poor town in Kentucky. It makes little sense to guard against the revelation that, as one would expect, there are no billionaires in Booneville at the cost of the utility of the rest of the dataset. Differential privacy does not differentiate between the <sup>93</sup> many possible types of revelations. It treats all as if they were equally meaningful, which leads to silly results and upside-down priorities.

### C. Tables

Part I demonstrated that differential privacy can perform fairly well when queries are asked to report counts, such as the numbers of people who have various characteristics. Suppose that, instead of querying the mean income, the data user submitted a query to create a histogram of income? With count queries, the addition or deletion of one individual changes only a single bucket in a histogram--and by only 1. Thus, the global sensitivity is 1 instead of \$1 billion.

Before we present the results, it is worth reflecting on the loss of utility that comes with the change of format. The accuracy of simple statistics from grouped histogram data is always compromised by the crudeness of the categories. Still, one might expect an improvement over the differential privacy responses for average income that we explored above.

Table 6 shows a hypothetical histogram for Booneville, Kentucky, and noise that we randomly selected from a Laplace distribution with  $\epsilon = 0.50$  (as before). This is just one realization of possible responses to the histogram query. In practice, the data user would see only the last column of the table. The shaded columns help us assess whether the last column is close enough for research purposes.

**Table 6--Example Responses to a Series of Count Queries about the Income of Booneville Residents <<unknown symbol>><sub>q</sub> = 0.5, Δf = 1**

Income Group	True Count	Noise (rounded to the closest integer)	Response (True Count + Noise)
\$0 to \$10 Thousand	11	2	13
\$10 Thousand to \$50 Thousand	40	7	47
\$50 Thousand to \$100 Thousand	7	2	5
\$100 Thousand to \$500 Thousand	1	4	3
\$500 Thousand to \$1 Million	0	5	5
\$1 Million to \$10 Million	0	0	0
\$10 Million to \$100 Million	0	3	3
\$100 Million to \$1 Billion	0	0	0
More than \$1 Billion	0	5	5

\*732 The unshaded response column reports that there are five individuals whose income is higher than \$1 billion and three individuals whose income is between \$10 million and \$100 million. Of course, we know that the maximum income of individuals in Booneville city is less than \$500,000, so this table steers researchers wildly off the mark.<sup>93</sup> Naturally, the negative values are par for the course.<sup>94</sup> They very slightly help balance out the bias from positive noise if the researcher decides to use the table to calculate a rough estimate of average income, but the correction is hardly worth the bother since an estimate of the average would be quite poor as it is. A researcher using only the responses above would conclude that the average income among Booneville residents is about \$44 million.<sup>95</sup>

Why does this table perform so poorly even though the table from Part I, reporting the number of millionaires in Washington, performed so well? Recall that the noise or, more precisely, the distribution that produces the noise, is independent from the true values in the original dataset. It is also independent from the size of the database. In both tables, the global sensitivity (Δf) is 1. However, when working with the number of Washington millionaires, noise in the range of 7 to 7 does not make much of a difference because the true response is over 200,000. Here, since the true answers are small (under 100), noise on the same scale greatly distorts the analysis.

Table 7 shows the Laplace distributions for tabular data, where Δf = 1. Each row displays the probability of observing noise values within the identified range for varying specifications of <<unknown symbol>>.

**\*733 Table 7--The Probability that Laplace Noise Will Be Selected from Specified Ranges, for Varying Selections of <<unknown symbol>> Δf = 1**

	0.001	0.01	0.10	0.25	0.50	ln(2)	1.00	ln(3)	5.00
±1	0.00	0.01	0.10	0.22	0.39	0.50	0.63	0.67	0.99
±2	0.00	0.02	0.18	0.39	0.63	0.75	0.86	0.89	1.00
±3	0.00	0.03	0.26	0.53	0.78	0.88	0.95	0.96	
±5	0.00	0.05	0.39	0.71	0.92	0.97	0.99	1.00	
±10	0.01	0.10	0.63	0.92	0.99	1.00	1.00		
±20	0.02	0.18	0.86	0.99	1.00				
±50	0.05	0.39	0.99	1.00					
±100	0.10	0.63	1.00						
±500	0.39	0.99							
±1000	0.63	1.00							
±5000	0.99								
±10000	1.00								

When  $\epsilon > 1$ , relatively little noise is added to the true answer. But, large  $\epsilon$  values open the system to risk of disclosure, and the risk is not managed in any thoughtful way. When  $\epsilon$  is as large as 5 or higher, the risk of disclosure is so great that the system cannot fairly be described as a privacy-protecting one. When  $\epsilon < 0.10$ , the noise generated could be  $\pm 100$ . Adding 100 or more to a query response might be just fine if the true response is in the order of 100,000 or more, but it causes chaos if the true answer is less than ten. Table 7 shows the distribution of noise added to count queries irrespective of the true answer. Once  $\epsilon$  is specified, the noise will be generated with the above stated probabilities.

Dwork defends this as a desirable feature since small databases leave the data subjects more vulnerable and thus require proportionally more protection than larger databases.<sup>96</sup> But this is not necessarily so. Suppose that Table 6, the representative example of a histogram query, reports the income not from the town of Booneville, but from a stratified random sample of 130 Americans. As long as the adversary does not have a way of knowing who was included in the random sample, this database would not require any more protective noise than a database containing the entire US population, yet **\*734** differential privacy methods would cause much more loss to its utility.<sup>97</sup>

Moreover, the noise distribution is not limited to count queries. This noise is added in all situations for which  $\Delta f = 1$ , even if the query demands a strict upper and lower bound for the true value. Consider the query, "What is the average income tax rate for Americans?" A person submitting the query would expect a reasonable response between 0% and 39.6% (the highest marginal tax rate), but Table 7 shows that for any  $\epsilon < 5.0$ , there is a high probability that the response will be negative or above 1, rendering it useless. This also poses a significant problem for statistical measures that must be interpreted within a bounded range, as we illustrate in the next example.

#### D. Correlations

Lest there be any doubt that differential privacy performs poorly under most typical research settings, consider its effects on correlation. Statistical research often explores the relationships between variables. Pearson's product-moment correlation, measuring the strength of the linear relationship between two variables, is one of the most basic and essential tools to understand how various forces and phenomena interact and operate on one another. Correlation ranges between  $[-1, 1]$  where -1 means that two variables have a perfectly negative relationship (an increase in X corresponds with a proportional decrease in Y), 0 means the two variables share no relationship (an increase in X sometimes corresponds with increases and sometimes decreases in Y), and 1 indicates a perfectly positive relationship (an increase in X corresponds with a proportional increase in Y). In this case, the function (correlation) has clear lower and upper bounds--a query on correlation will always come out between -1 and 1.

Suppose the Department of Education is preparing a database query system based on a national longitudinal study on the relationship between education and income. Among other things, the database contains information on each data subject's highest educational attainment (measured in years of qualified schooling) and annual income. What happens when the Department of Education adopts differential privacy and applies Laplace noise to a query **\*735** requesting the correlation between educational attainment and income?

Let us work through the usual steps:

1. Select  $\epsilon$

In this example, let us explore what happens to the query response under a range of  $\epsilon$  running from 0.01 (relatively privacy protective) to 10.0 (quite lax). As before, we will assume a single query of the database to avoid the need to add more noise for serial queries.

2. Compute the Response to the Query Using the Original Data



The relationship between education and income is strong. Expected earnings increase in lockstep as a person moves from high school to college to masters, doctoral, or professional degrees.<sup>98</sup> Assume for this exercise that the education and income data in the Department of Education's database produce a correlation coefficient of 0.45.

3. Compute the Global Sensitivity  $\Delta f$  for the Query

The global sensitivity requires the data curator to anticipate the greatest difference that the addition or subtraction of a single data point can make to a similar query on the same variable for any possible database--not just the database that the curator is preparing for public research.<sup>99</sup>

For a very small sample, the addition (or subtraction) of a single data subject can change the correlation coefficient of two variables from perfectly positive correlation to a strong negative correlation, or vice versa--a change of nearly 2. To see how, imagine a database with just two people. Person A has had fewer than 8 years of formal education (no high school) and has an annual income of \$52,000. Person B has a professional degree and earns \$70,000 each year. For this small set of data, correlation between education and income will be 1: the more education, the more income. Now, imagine what happens when we add Person C to the dataset. Person C also has no formal education, but has an income of \$1 million. With these three data points, the correlation between income and education can \*736 fall below 0. After adding Person C, it looks like on balance, less education will tend to increase income.

We could construct a similar illustration where a correlation of +1 is converted to 1 (or something infinitely close) with the addition of 1 new data point, so we are working with  $\Delta f = 2$ .

4. Generate a Random Value (Noise) from a Laplace Distribution with Mean = 0 and Scale Parameter  $\Delta f$

Next we randomly draw noise from the Laplace distribution determined by the values of global sensitivity and  $\Delta f$ . This is where the process takes a turn for the worst.

Correlation takes the range from -1 to 1. Output outside of that range would be meaningless, and small changes within the range can have a great effect on the researcher's interpretation. Table 8 reports the probability that the noise added to the true answer will be no higher than 1, and no lower than -1 under varying selections of  $\Delta f$ .

**Table 8--The Probability that Laplace Noise Will Fall Within [-1, 1] for Varying Selections of  $\Delta f$**

$\Delta f$	Probability Noise Is in the Range [-1, 1]
0.01	0.004988
0.10	0.048771
0.20	0.095163
0.50	0.221199
1.00	0.393469
2.00	0.632121
5.00	0.917915
10.00	0.993262

For small, privacy-protecting levels of  $\Delta f$  (< 0.50), the noise added to the true answer is very likely to be so large that the query system's response will be nonsense. If the data curator selects  $\Delta f \geq 5$ , there is a decent chance the reported correlation will be within the range, but of course it is also very likely to misstate the relationship between the variables (and to say that two factors that are positively correlated are negatively correlated, or vice versa).

\*737 Figure 4 shows what the distribution of responses would be, assuming that the true answer (the actual correlation) is zero and  $\sigma = 0.50$ . The dotted lines show the acceptable response range  $[-1, 1]$ . The figure illustrates that the great majority of responses would fall outside the acceptable range for correlation rendering the response completely meaningless to the user. Many of the responses within the dotted lines would be very misleading to the researchers and to the relying public.

**Figure 4--Distribution of Responses to a Query for Correlation Where the True Answer is 0  $\sigma = 0.5$ ,  $\Delta f = 2$**

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

With noise like this, differential privacy simply cannot provide a workable solution for analyses of correlations or of any statistical measure with a strict upper and lower bound.

The examples worked through in this Part should give a sense of differential privacy's serious practical limitations. While differential privacy is a technical standard, the problems that it would cause if adopted broadly would be profound, wide reaching, and devastating to research. Nevertheless, policymakers and privacy \*738 scholars are embracing differential privacy with increasing enthusiasm.<sup>100</sup> This enthusiasm must be tempered. The proponents of differential privacy have oversold its usefulness. Realistically, the future of data privacy will rely on differential privacy only in very narrow circumstances or only if differential privacy is modified to the point of being unrecognizable to its creators.

### III. The Golden Hammer

The proponents of differential privacy have embraced the law of the instrument: When you have a hammer, everything looks like a nail. The developers of differential privacy have insisted that it is a full-service tool that will free research from the perils of privacy risk in every context. As Cynthia Dwork and her collaborators say, apply differential privacy “and never look back.”<sup>101</sup>

Policymakers and legal scholars are ready to adopt differential privacy as a--or even the--best practice, though their enthusiasm reveals a lack of understanding about what differential privacy would do to data research.<sup>102</sup> In one case, legal scholars jumped to the conclusion that Facebook employs differential privacy when it is very likely using a different noise-adding technique.<sup>103</sup> This is a variation on the law of the instrument: When you like hammers, every tool looks like one.

In this Part, we will explore why differential privacy has suddenly gained the attention and trust of legal scholars and policymakers. Without exception, the enthusiasm for differential privacy stems from misinformed understanding of how the standard works. This Part also explores instances where differential privacy will likely work well and where it will likely not.

#### \*739 A. Misinformed Exuberance

The examples worked through in Part II showed that differential privacy has serious practical limitations. Somehow these problems have escaped the notice of many scholars and journalists, even when the drawbacks are right under their noses.

Consider this excerpt from a Scientific American article:

Suppose the true answer to [a query] is 157. The differentially private algorithm will “add noise” to the true answer; that is, before returning an answer, it will add or subtract from 157 some number, chosen randomly according to a predetermined set of probabilities. Thus, it might return 157, but it also might return 153, 159 or even 292. The person who asked the question knows which probability distribution the algorithm is using, so she has a rough idea of how much the true answer has likely been distorted (otherwise the answer the algorithm spat out would be completely useless to her). However, she doesn't know which random number the algorithm actually added.<sup>104</sup>

This is a typical explanation and endorsement of differential privacy and it makes an equally typical mistake. The author starts with an assumption that contorts the rest of her analysis. The key here is that the reader already knows what the true answer is--157. It is only if the reader already knows the answer that a response like "159 or even 292" can seem useful. But how would the hypothetical researcher, who must operate in ignorance of the true answer, react to a response of "159 or even 292?"

Now consider how the query response in this hypothetical could be meaningful. First, the response might be useful if the selected <<unknown symbol>> is large, so that the magnitude of the noise is very likely to be small. But the author says the response could very well be 292. If the noise added spans a range of 150, <<unknown symbol>> in this case cannot be small. We can rule out this possibility.

The second possibility is that a span of 150 might still be small relative to the sort of numbers the researcher was expecting to observe. For example, if the questioner had asked a database containing information on the entire US population to return the number of people who live in particular town in order to understand whether the town is big or small, then a response within 150 of the true value sheds some light. As we have said before, count queries that happen to have very large values are suitable for differential privacy techniques.<sup>105</sup> However, these are unusual conditions. For most researchers, an answer that is likely to be 150 away from the true answer, and that allows them only to conclude things like "this is large-ish" or "this is probably small" will not be good enough. After \*740 all, from the perspective of a researcher who does not know the true answer, a query response of "292" with a margin of error in excess of 150 would have to consider that the true answer might be 442, and that is quite far off from the true answer, which we know to be 157.

The Scientific American journalist assumed that the questioner already knew the true answer, or, at least, has a good sense of its ballpark.<sup>106</sup> The experience of a researcher who already knows the answer makes a lousy gauge for the utility of a query system. Instead, we should be concerned about the researchers who potentially do not know what the approximate true answer is. After all, if the researcher knew the approximate answer, he would have little reason to use a query system that adds noise. Scientific American thus relays some of the misplaced confidence of the developers of differential privacy.

We take our next example from a Microsoft whitepaper titled Differential Privacy for Everyone.<sup>107</sup>

A researcher wants to test whether a particular disease is more likely to manifest in people who have lived in certain regions. She connects to her hospital's query system that has differential privacy guards in place. The researcher makes a series of queries on the number of patients with the disease who have lived in each of the towns in the suspected region. Suppose that some of the towns have a large number of people with the disease, some towns have no people with the disease, and one town, Smallville, has a single case. If the query system were to report the true answers to the researcher, the patient (Bob) in Smallville may be at risk. For example, if he had very recently moved to the researcher's hometown, and the researcher knows he is from Smallville, she might be able to put together that he has the disease. The Microsoft whitepaper explains:

To avoid this situation, the [query system] will introduce a random but small level of inaccuracy, or distortion, into the results it serves to the researcher. . . .

. . . .

Thus, the answers reported by the [query system] are accurate enough that they provide valuable information to the researcher, but inaccurate enough that the researcher cannot know if Bob's name is or is not in the database.<sup>108</sup>

The conclusions that Microsoft urges us to draw are speculative, to say the least. There is simply no guarantee that the responses from the query system would lead the researcher to the correct approximate understanding about where the cases of the \*741 disease do and do not come from. Whether the responses are only “slightly larger or smaller” will depend entirely on the data curator's specification of <<unknown symbol>> and the total number of queries.<sup>109</sup>

For good measure, let us quickly work through the hypothetical selecting a relatively liberal value for <<unknown symbol>> (that is, a less privacy-protecting choice). Suppose <<unknown symbol>> = ln(3), which is approximately 1.0986. Assume also that the curator of the database has determined that a total of 1000 simple count queries can be issued to the database. Allowing a range of queries would require us to add more noise, so this is a realistic lower bound in terms of the distortion of results.

With <<unknown symbol>> = 1.0986 and <<unknown symbol>> = 1000, we must use << unknown symbol>>= (1.0986/1000) for each individual query. As with all count queries, the most a single individual can influence a count query is by 1, so <<unknown symbol>> = 1.<sup>110</sup>

What happens when the researcher queries the system “For each town located in the suspected regions, what is the number of patients with the disease?” Table 9 reports the likelihood that the noise added to each town's response will be within a particular range.

**Table 9--Distribution of Laplace Noise Within Specified Ranges <<unknown symbol>> = ln (3)/1000, Δf= 1**

Noise Range	Probability
±1	0.00
±5	0.01
±10	0.01
±50	0.05
±100	0.10
±500	0.42
±1000	0.67
±10000	1.00

So, for Smallville, there is a very high chance--16%--that the response will exceed 1000, even though we know the true answer is 1. There is also a very high chance--again, 16%--that the response will be less than 1000.

\*742 Now consider one of the towns “where there are a significant number of individuals” with the disease. Suppose the number of individuals with the disease is about 100. The response has a 45% chance of having a zero or negative value. Even if the number of individuals with the disease in this town is 1000, the probability of observing a negative value response is greater than 16%. Therefore, it is not obvious at all that a faithful use of differential privacy will provide the researcher with meaningful answers from which she could infer that eight towns had a number of people with the disease, and Smallville had either a small number or 0.

To drive this point home, Table 10 provides just one realization, selected randomly from the Laplace noise distribution, for the eight towns and Smallville.

**Table 10--Example of Noise-Added Responses to the Smallville Hypothetical << unknown symbol>> = ln (3)/1000, Δf= 1**

Town	True Answer	Noise	Response
1	105	2893.9	2998.9

2	80	2840.6	2760.6
3	92	848.6	940.6
4	100	4099.3	4199.3
5	125	2145.4	2270.4
6	103	1607.8	1504.8
7	99	814.6	715.6
8	85	191.3	276.3
Smallville	1	817.3	818.3

The researcher, who sees only the unshaded last column, would be hard-pressed to say anything about the relative prevalence of the disease in these nine towns. The best the researcher could do is conclude that, knowing the value of <<unknown symbol>>, the true responses were not large enough to overpower the magnitude of the noise that had to be added to maintain differential privacy. The researcher could conclude that none of the towns had tens of thousands of cases of the disease, but she could not confidently say anything more specific than that.

The only practical application of this sort is in response to queries involving common diseases like the flu that occur in the tens of thousands across the subpopulations of interest. For a rare form of cancer, answers drawn from the differential privacy parameters we set will be useless, or worse than useless.<sup>111</sup>

**\*743** The curator could try to set the parameters differently from ours in order to squeeze some more utility out of the system. The curator could, for example, decide that the system will only respond to a small number of queries so that the <<unknown symbol>> for each query could be larger. But by reducing the number of queries, the curator reduces the overall value of the query system.<sup>112</sup>

The Microsoft authors' reassurance that "the answers reported by the DP guard are accurate enough that they provide valuable information to the researcher" is thoroughly unwarranted. Reassurances of this sort mislead lay audiences into the optimistic impression that differential privacy preserves data utility better than it does.

By working with examples where they already know the true answer, the proponents of differential privacy have given the impression that the standard is more useful and viable than it really is. Erica Klarreich, the author of the Scientific American article, advances the following illustration:

To see what kind of distribution will ensure differential privacy, imagine that a prying questioner is trying to find out whether I am in a database. He asks, "How many people named Erica Klarreich are in the database?" Let's say he gets an answer of 100. Because Erica Klarreich is such a rare name, the questioner knows that the true answer is almost certainly either 0 or 1, leaving two possibilities:

(a) The answer is 0 and the algorithm added 100 in noise; or

(b) The answer is 1 and the algorithm added 99 in noise.

To preserve my privacy, the probability of picking 99 or 100 must be almost exactly the same; then the questioner will be unable to distinguish meaningfully between the two possibilities.<sup>113</sup>

The assumption that "the questioner knows that the true answer is almost certainly either 0 or 1" turns out to be critical to understanding whether differential privacy is striking the right balance between privacy and utility. We might be satisfied that

this intrusive data user must ignore the response to his query because, in the trade-off between his curiosity and Erica Klarreich's privacy, the better interest prevailed.

\*744 But what if the questioner does not know the true answer must be 0 or 1? Instead of “How many people named Erica Klarreich are in the database?” what if the query was “How many people died of postoperative infections last month at this hospital?” Now, when the user receives the response “100,” he will either naively assume that the hospital must have terrible sanitary conditions, or, if he is a sophisticated user, he would know to ignore the results since the probability distribution of the noise is in the order of  $\pm 100$ .

Thus, although we changed nothing about the differential privacy mechanism (altering only the intent of the data user, who in this case is not malicious), a result of “100” to a query whose true result is 0 or 1 is no longer satisfactory. After all, if the true answer is 0, we would not want the data user to worry about the conditions of the hospital. But if the true result were close to 100, we would want the researcher to worry. If a hospital were to create a publicly available query system, it would have to anticipate both types of queries—that is, both the intrusive “how many people named Erica Klarreich” query and the postoperative infections query.

The best way to avoid the absurdities is for data curators to ensure that the magnitude of the noise added to a query is comparable to the true answer. But context-driven addition of noise would violate the basic tenets of differential privacy.<sup>114</sup> To satisfy differential privacy, the noise must be independent, not only of the true answer, but also the size of the database.<sup>115</sup> Legal scholars and policymakers have overlooked this drawback.

## B. Willful Blindness to Context

One of differential privacy's strongest and most attractive claims is that it can—and in fact must—be applied without considering the specifics of the queried database.<sup>116</sup> But as we saw with the average income example, the blindness to context has harsh consequences. If databases must protect Bill Gates, George Soros, and other highly unusual individuals, then the curator has only two realistic options: give up on utility, or give up on privacy.

When scholars and journalists provide examples of differential privacy in action, they invariably use tables of counts to show how it works.<sup>117</sup> But statistical research often involves the analysis of numerical data. Our examples show that differential privacy is \*745 unlikely to permit meaningful results to queries for averages and correlations unless the data curator selects a very high <<unknown symbol>>, but in that case, the curator has abdicated his chance to protect privacy.

The natural desire to avoid absurd results has led some supporters of differential privacy to mischaracterize, possibly even misunderstand, what differential privacy demands and to insist that the characteristics of a database, or the answer to a particular query, has some influence over the noise that is added.<sup>118</sup> For example, Felix Wu describes differential privacy as follows:

The amount of noise depends on the extent to which the answer to the question changes when any one individual's data changes. Thus, asking about an attribute of a single individual results in a very noisy answer, because the true answer could change completely if that individual's information changed. In this case, the answer given is designed to be so noisy that it is essentially random and meaningless. Asking for an aggregate statistic about a large population, on the other hand, results in an answer with little noise, one which is relatively close to the true answer.<sup>119</sup>

Contrary to Wu's assertion, differential privacy noise is not a function of the breadth of the query. Because the noise is based on global sensitivity, for all databases that could possibly exist, the noise added to any particular query response must be the same whether the query involves a single person or a million. When it comes to counts and tabular data, the noise added to a query on a large number of people might be less distorting than noise of the same size added to a query on a small number of subjects. But, with other analyses (like correlation), the distortions will be equally severe no matter the  $n$ .<sup>120</sup> Lest there be any doubt, Dwork herself has recently insisted, "Our expected error magnitude is constant, independent of  $n$  [the number of data subjects responsive to a query]."<sup>121</sup>

A white paper from Microsoft's differential privacy research team makes a similar error.<sup>122</sup> It states:

Distortion is introduced into the answers a posteriori. That is, the DP guard gets answers based on pristine data, and then mathematically decides the right amount of distortion that needs to be introduced, based on the type of question that was asked, on the size of the database itself, how much its data changes on a regular basis, etc.<sup>123</sup>

Wu and the authors of the Microsoft paper are unwittingly rewriting how differential privacy works. Wu implies that what matters is the influence that a particular piece of information can have on the particular query that has been submitted. This would be a \*746 fabulous improvement for preserving the utility of a dataset, but it cannot promise differential privacy because a series of queries could reveal changes in the magnitude of the noise that would reveal information about the underlying values.<sup>124</sup> Thus, the technical literature on differential privacy has consistently maintained that the magnitude of the noise must be independent of the size of the data set, the magnitude of the true answer, and the type of query (except in assessing << unknown symbol >>, which requires an assessment of all possible query responses across the universe of possible datasets).<sup>125</sup>

Finally, Ed Felten, Chief Technologist of the Federal Trade Commission, describes differential privacy as if it curbs the amount of error around a particular response. He uses the following example:

Let's say [an adversary's] best guess, based on all of the available medical science and statistics about the population generally, is that there is a 2% chance that I have diabetes. Now if we give the [adversary] controlled access to my doctor's database, via a method that guarantees differential privacy at the 0.01% level, then the analyst might be able to adjust his estimate of the odds that I have diabetes-but only by a tiny bit. His best estimate of the odds that I am diabetic, which was originally 2%, might go as low as 1.9998% or as high as 2.0002%. The tiny difference of 0.0002% is too small to worry about.

That's differential privacy.<sup>126</sup>

This is not differential privacy at all. An adversary could query the database for the proportion of patients in the doctor's database who have diabetes. This ratio could significantly improve the adversary's guess for Ed Felten's likelihood of having diabetes. This is especially true if the doctor's practice is large enough so that the noise does not drown out the true response.<sup>127</sup> It is also especially true if Ed Felten's doctor specializes in the treatment of diabetics. So Felten's claim can only be correct if we assume that the proportion of individuals with diabetes in his doctor's practice happens to be 2%, just like the general public.

Felten's example illustrates the sort of willful blindness to context that comes from a threat model orientation. By focusing exclusively on the adversary, Felten fails to see the consequences to legitimate research. In a realistic scenario, the number of patients in \*747 the doctor's database is likely to be a few thousand.<sup>128</sup> A query system using << unknown symbol >> = 0.0001

would have to add tremendous noise to each response.<sup>129</sup> The answers are unlikely to be anywhere close to the true value--whether the legitimate user queries the doctor's database for a count of the number of patients with diabetes or asks point blank "Does Ed Felten have diabetes?" The consequences to research are an afterthought for the proponents of differential privacy.

The legal scholars and policymakers who endorse differential privacy do so only when (and because) they think it works differently than it really does.<sup>130</sup> Differential privacy eschews a nuanced approach that takes into account the variety of disclosures relatively likely to occur, the underlying data, and the specifics of a particular query. This "one size fits all" solution has exactly the problems that one would expect from a nonnuanced rule. It behaves like Procrustes's bed, cutting off some of the most useful applications of a query system without reflection on the costs.

### C. Expansive Definitions of Privacy

Differential privacy is motivated by statistician Tore Dalenius's definition of disclosure, which identifies any new revelation that can be facilitated by a research database as a reduction of privacy.<sup>131</sup> As Dalenius well knew, eliminating this type of disclosure is not only impossible, it is not even the right goal.<sup>132</sup> Differential privacy makes no differentiation between the types of auxiliary information that an intruder may or may not have. Because it remains agnostic to these types of considerations, the assumptions about what an attacker might know are unrealistic and too demanding. In order to make differential privacy protections manageable, data curators will be tempted to choose a large value for  $\epsilon$  or to relax the standards in some other way. But this will relax the privacy protections in a thoughtless way, divorced from context, and thus runs the risk of exposing a few data subjects to unnecessary risks. Embracing too expansive a definition of disclosure creates the danger that curators will deviate from the standard without assessing which disclosures are important (e.g., an increased chance of inferring that Bob has HIV) and which are not (e.g., a decreased chance of inferring that Bob is not a billionaire).

The expansiveness of differential privacy comes from its anticipation of all databases in the universe. Differential privacy defines privacy breach as the gap in probabilities of observing a particular response, not for the particular database in use, but for all possible datasets  $X$  and  $X^*$  that differ on, at most, one row.<sup>133</sup> This is why we have to consider George Soros's income when we are dealing with the income of the citizens of Booneville.

The rationale for this requirement comes from the fact that we not only have to provide protection for the citizens of Booneville, but we must also prevent the response from revealing that someone is not a citizen of Booneville. This is true even if it is generally known that George Soros is not a citizen of Booneville and that Booneville does not tend to attract people with wealth. Thus, what may have looked like an advantage of differential privacy--that it requires no assumptions about what adversaries already know--is actually a stumbling block. It causes differential privacy to obliterate accurate responses with noise. By calibrating to the most extreme case (i.e., George Soros), differential privacy protects everyone, but only at significant cost to research.

This explains why differential privacy seems to work pretty well for some counts of individuals but not so well for other variables. For counts, every person exerts the same level of influence and  $\Delta f = 1$  regardless of who is or is not included in the database.<sup>134</sup> But for other variables, such as income, the influence exerted by an outlier is very different than that exerted by nearly every other entry. Attempting to protect George Soros's income information adds so much noise that it overwhelms the information about the income of the average citizen (from Booneville or any other city). Dwork obliquely acknowledges as much when she says, "Our techniques work best - i.e., introduce the least noise - when  $\Delta f$  is small."<sup>135</sup> What is left unsaid is that when  $\Delta f$  is very large, differential privacy simply breaks down.

Comparing two databases that differ in one record from the universe of all databases leads to the popularized claim of differential privacy "that it protects against attackers who know all but one record."<sup>136</sup> The negative consequences of this requirement are less well known. Differential privacy provides protection in anticipation of the worst-case scenario, which is admirable, but impractical. We could build every building as if it were Fort Knox--but at what cost?



**D. Multiple Queries Multiply the Problems**

The effect of differential privacy protections on each query is cumulative.<sup>137</sup> This is one of the least discussed factors in the implementation of differential privacy. Any reasonably sized database--such as that of a healthcare provider--is likely to be queried thousands of times. For databases released by government agencies, such as the Census Bureau, the number of queries could easily reach the millions. This is likely true for large databases held by Facebook, Google, and others.<sup>138</sup> If the curator provides responses to a set of  $m$  separate queries with privacy parameter  $\epsilon$ , then the global privacy measure for the database is  $m\epsilon$ , and thus the differential privacy risk  $\epsilon$ .<sup>139</sup> That is, the differential privacy standard is the sum of all the query epsilons.<sup>140</sup> If the curator wants to keep the global  $\epsilon$  under 10, he would have to set either  $\epsilon$  (the  $\epsilon$  for each query) or  $m$  (the number of queries) to be quite small. In either case, this severely limits the usefulness of the database. Neither is desirable.

A majority of statistical analyses, such as hypothesis testing, relies on at least the mean and variance--or in the case of multiple variables, the means and the correlations. When every quantity is a "noise-added" response, the effects of large noise-addition can lead to meaningless, or even dangerous, conclusions.

**\*750 E. At the Same Time, Limited Definitions of Privacy**

Differential privacy ensures that an individual's inclusion or exclusion from the dataset does not change the probability of receiving a particular query response by too much, but meeting this standard does not necessarily guarantee privacy in the conventional sense.

First, differential privacy leaves the designation of  $\epsilon$  to the discretion of the data curator.<sup>141</sup> If the curator is committed both to differential privacy and to maintaining the utility of the data query system, he will be tempted to select a large  $\epsilon$  and to allow a large number of queries. If the curator selects a large  $\epsilon$ , the standard will be so relaxed that the benefits of differential privacy are wasted. For example, suppose the curator selects  $\epsilon = 10$ . 10 sounds like a reasonable enough number, but the privacy standard is actually  $\epsilon$ . So when  $\epsilon = 10$ , the ratio of probabilities for a result with and without the inclusion of an individual can be over 22,000. The ratio just need be less than  $e^{10}$  (about 22,026.3).<sup>142</sup> With probabilities this different, the curator would have more luck protecting the privacy of the data subjects by adding random noise selected within some context-appropriate bounded range. If the  $\epsilon$  is large, the protections offered are hardly worth the effort. The nature of exponents is such that small differences in  $\epsilon$  cause very large differences in privacy protection. Table 11 shows the powers of  $e$ .

**Table 11--Differential Privacy Standards (Ratio of Probabilities) for Varying Selections of  $\epsilon$**

$\epsilon$	$\epsilon$	$\epsilon$	$\epsilon$
0.01	1.01	$\ln(3)$	3.00
0.05	1.05	2	7.39
0.10	1.11	5	148.41
0.25	1.28	10	22,026.47
0.50	1.65	25	$7.2 \times 10^{10}$
$\ln(2)$	2.00	50	$5.18 \times 10^{21}$
1.00	2.72	100	$2.68 \times 10^{43}$

Let us work through a quick example of what happens when the curator decides to answer one thousand queries from the \*751 Booneville City database (which may contain, in addition to income, a lot of other information about the citizens of

Booneville). For a single query, we observe that the probability of observing a response within  $\pm \$1$  million is approximately 3% (and 97% of the time it was higher than this range). To be equitable, we assume that every query will be answered with the same level of privacy (assuring both equally accurate responses to all queries and equal privacy for all citizens) resulting in  $\llbracket \text{equation} \rrbracket$ . This means that the noise added would increase thousandfold.<sup>143</sup> With one thousand queries, the observations for average income over a small town would be laughably wrong. The query system would provide responses within \$1 billion of the true answer about 3% of the time. The rest of the time (the remaining 97%), the response will be greater than \$1 billion or less than negative \$1 billion.

Dwork occasionally underplays the importance of the selection of  $\llbracket \text{unknown symbol} \rrbracket$  to guard against potential privacy-invading uses. She states “if the [differentially private] database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of any individual's data in the database will not significantly affect his or her chance of receiving coverage.”<sup>144</sup> But with a high enough  $\llbracket \text{unknown symbol} \rrbracket$ , an insurance adjuster could take advantage of the lax standard. For example, suppose the adjuster asks, “Does Jeff Jones have a congenital heart disease?” and  $\llbracket \text{unknown symbol} \rrbracket$  is set to  $\ln(2)$ . This means that the ratio of probabilities that the database will give a particular response equals 2. Thus, if Jeff Jones were to have the disease, it is twice as likely to observe a response that he has the diseases compared to the response that he does not have the disease.<sup>145</sup> So when they receive a positive response, the insurance company may want to play the odds and decline coverage.

The effects are worse for clusters of individuals. Consider an insurance company employee who issues the query, “How many individuals in the Jones family of 5 have a congenital heart disease?” Assuming one or more of the individuals in this family does have the congenital heart disease, the probability of a response indicating that \*752 one or more individuals in this family has the disease is 32 times ( $2^5$ ) more likely than a negative response because differential privacy ensures only that each marginal individual contribute no more than a doubling of the probability. For five individuals in a row, the ratio would double five times. Now, the insurance adjuster is very likely to decline coverage for the Jones family since the chance that all of them don't have heart disease may be a paltry  $1/33$ .<sup>146</sup>

## F. Difficult Application

Because differential privacy techniques are agnostic to the specific underlying database, one might get the impression that they are easy to implement. This is not the case.

In order to create the appropriate Laplace noise distribution, the data curator must identify and assess the global sensitivity ( $\Delta f$ ) for every type of allowable query.<sup>147</sup> For some statistics, such as counts, sums, and mean, the analysis is straightforward. For most tabular data,  $\Delta f = 1$ .<sup>148</sup> Sums and means require the curator to know the largest values over the entire world's population for each variable, but as long as they have access to some reliable descriptive statistics<sup>149</sup>, this is usually not too hard.

For analyses involving more complicated statistics, determining global sensitivity is not an easy task. Consider the illustration in which a user queries a database for the average income of residents in Booneville, Kentucky. In order to compute  $\Delta f$ , the data steward will have to guess the income of the world's highest-paid person. Error has serious consequences: under-specifying  $\Delta f$  would mean that differential privacy is not actually satisfied, but over-specifying  $\Delta f$  will further degrade the quality of the output. Statistical analysis often involves estimates of important statistical \*753 relationships between numerical variables such as variance, regression coefficients, coefficient of determination, or eigen-values. For these types of queries, determining global sensitivity will be very challenging. Correctly choosing global sensitivity has drastic consequences to utility--as we saw with the correlation example in Part II.

Considering all of these limitations together, we must circumscribe the practical applications for pure differential privacy to the situations in which count queries have true answers that are very large. Unless we alter the core purposes and definitions of differential privacy, statisticians and policymakers should ignore the hype.

## Conclusions

Differential privacy faces a hard choice. It must either recede into the ash heap of theory or surrender its claim to uniqueness and supremacy. In its pure form, differential privacy has no chance of broad application. However, recent research by its proponents shows a willingness to relax the differential privacy standard in order to complex queries. Two such relaxations are often used.

The first, proposed by Dwork herself, requires that the probability of seeing a response with a particular subject remain within some factor of the probability of the same response without that subject plus some extra allowance.<sup>150</sup> The problem with this modification is that there is no upper bound on the actual privacy afforded by this standard.<sup>151</sup> In some situations, this allowance may be appropriate, but it would require the judgment of a privacy expert based on context--the very thing differential privacy had sought to avoid.

Ashwin Machanavajjhala developed another alternative for the US Census Bureau's On the Map application.<sup>152</sup> This relaxation of differential privacy allows curators to satisfy a modified differential privacy standard while usually meeting strict differential privacy. For some predesignated percentage of responses, the differential privacy \*754 standard can be broken.<sup>153</sup> This relaxation also undermines the promise of privacy.<sup>154</sup> In the situations where differential privacy is not satisfied, there is no upper bound on the risk of disclosing sensitive information to a malicious user. However, this may be fine if the curator crafts the deviations in a thoughtful way. Nonetheless, the data curator would need to resort to judgment and context.

This progression by the differential privacy researchers to a relaxed form is odd, given their view that historical definitions of privacy in the statistical literature lack rigor. The differential privacy community roundly dismisses traditional mechanisms for not offering strong privacy guarantees,<sup>155</sup> but the old methods will often satisfy the proposed relaxed forms of differential privacy as On the Map clearly illustrates.<sup>156</sup>

As differential privacy experts grapple with the messy problems of creating a system that gives researchers meaningful responses, while also providing meaningful disclosure prevention--albeit not differential privacy--they have come back to earth and rejoined the rest of the disclosure risk researchers who toil with the tension between utility and privacy.<sup>157</sup> In its strictest form, differential privacy is a farce. In its most relaxed form, it is no different, and no better, than other methods.<sup>158</sup>

Legal scholars and policymakers should resist the temptation to see differential privacy as a panacea, and to reject old disclosure prevention methods as inadequate. Adopting differential privacy as a regulatory best practice or mandate would be the end of research as we know it. The answers to basic statistical questions--averages and correlations--would be gibberish, and the standard would be very difficult to apply to regression and other complex analyses. \*755 Differential privacy would also forbid public microdata releases--a valuable public information resource.<sup>159</sup> Lest we end up in a land with a negative population of 30 foot-tall people earning an average income of \$23.8 million per year, the legal and policy community must curb its enthusiasm for this trendy theory.

## Footnotes

<sup>a1</sup> Associate Professor of Law, University of Arizona, James E. Rogers College of Law; J.D., Yale Law School; B.S., Yale College.

- aa1 Gatton Research Professor, University of Kentucky, Gatton College of Business & Economics; Ph.D., Texas A&M University; M.B.A., Sam Houston State University; B.Sc. University of Madras, India.
- aaa1 Ardmore Chair, Oklahoma State University, Spears School of Business; Ph.D., Texas A&M University; B.E., University of Madras, India.
- 1 See Eastern Equine Encephalitis, Centers for Disease Control & Prevention, <http://www.cdc.gov/EasternEquineEncephalitis/index.html> (last updated Aug. 16, 2010).
- 2 See Raghav Bhaskar et al., Noiseless Database Privacy, in *Advances in Cryptology - ASIACRYPT 2011: 17th International Conference on the Theory and Application of Cryptology and Information Security* 215, 215 (Dong Hoon Lee & Xiaoyun Wang eds., 2011); Samuel Greengard, Privacy Matters, 51 *Comm'n's of the ACM*, Sept. 2008, at 17, 18; Graham Cormode, Individual Privacy vs Population Privacy: Learning to Attack Anonymization, in *KDD'11 Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1253, 1253 (2011). But see Fida K. Dankar & Khaled El Emam, Practicing Differential Privacy in Health Care: A Review, 6 *Transactions on Data Privacy* 35, 51-60 (2013) (noting theoretical limitations that differential privacy must address before it can be widely adopted for health care research).
- 3 This is an actual instantiation of the differential privacy technique. The noise in this exercise was randomly drawn after setting  $\ll \text{unknown symbol} \gg = \ln(3)$  and allowing for 1,000 queries to the database. For a description of the technique, see *infra* Part I.B.
- 4 See Microsoft, *Differential Privacy for Everyone* 4-5 (2012), available at <http://www.microsoft.com/en-us/download/details.aspx?id=35409> (“Thus, instead of reporting one case for Smallville, the [query system] may report any number close to one. It could be zero, or (yes, this would be a valid noisy response when using DP), or even 1.”).
- 5 See Bhaskar et al., *supra* note 2, at 215; Cormode, *supra* note 2, at 1253-54; Greengard, *supra* note 2, at 18.
- 6 Google Scholar has indexed over 2,500 articles on the topic. Google Scholar, [www.scholar.google.com](http://www.scholar.google.com) (last visited Apr. 12, 2014) (describing a search for “Differential Privacy”).
- 7 Erica Klarreich, Privacy By the Numbers: A New Approach to Safeguarding Data, *Sci. Am.* (Dec. 31, 2012), <http://www.scientificamerican.com/article/privacy-by-the-numbers-a-new-approach-to-safeguarding-data>.
- 8 Greengard, *supra* note 2, at 18.
- 9 Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 *U. Colo. L. Rev.* 1117, 1139-40 (2013).
- 10 Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 *UCLA L. Rev.* 1701, 1756 (2010). Ohm acknowledges that differential privacy techniques add significant administration costs, and also risks denying the researcher an opportunity to mine the raw data freely to find useful patterns. *Id.* These are external critiques. Ohm does not present the internal critique of differential privacy theory that we develop here. See *id.*
- 11 Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 *N.C. L. Rev.* 1417, 1452-54 (2012).
- 12 Ed Felten, What Does it Mean to Preserve Privacy?, *Tech@FTC* (May 15, 2012, 4:47 PM), <http://techatftc.wordpress.com/2012/05/15/what-does-it-mean-to-preserve-privacy>.
- 13 See Frank D. McSherry, Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis, in *SIGMOD'09: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data* 19, 25 (2009); Klarreich, *supra* note 7.
- 14 See Bhaskar et al., *supra* note 2, at 215-16; Cynthia Dwork & Adam Smith, Differential Privacy for Statistics: What We Know and What We Want to Learn, 1 *J. Privacy & Confidentiality* 135, 139 (2009).
- 15 See George T. Duncan & Sumitra Mukherjee, Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks through Additive Noise, 95 *J. of the Am. Stat. Ass'n* 720, 720 (2000); Krishnamurty Muralidhar et al., A General Additive Data Perturbation Method for Database Security, 45 *Mgmt. Sci.* 1399, 1399-1401 (1999); Krishnamurty Muralidhar & Rathindra Sarathy, Data Shuffling--A New Masking Approach for Numerical Data, 52 *Mgmt. Sci.* 658, 658-59 (2006) [hereinafter Muralidhar & Sarathy, *Data Shuffling*]; Rathindra Sarathy et al., Perturbing Nonnormal Confidential Attributes: The Copula Approach, 48 *Mgmt.*

Sci. 1613, 1613-14 (2002); Mario Trottni et al., Maintaining Tail Dependence in Data Shuffling Using t Copula, 81 Stat. & Probability Letters 420, 420 (2011).

- 16 “Statistical offices carefully scrutinize their publications to insure that there is no disclosure, i.e., disclosure of information about individual respondents. This task has never been easy or straightforward.” I. P. Fellegi, On the Question of Statistical Confidentiality, 67 J. Am. Stat. Ass'n 7, 7 (1972).
- 17 These two popular forms do not exhaust the possibilities for data release, of course. Sometimes government agencies release summary information, such as a table, taken from more detailed data. These releases are neither microdata nor interactive data. See Jacob S. Siegel, Applied Demography: Applications to Business, Government, Law and Public Policy 175 (2002).
- 18 One popular form of microdata release is the “de-identified” public database. De-identification involves the removal of all personally identifiable information and, sometimes, the removal of other categories of information that can identify a person in combination. HIPAA, for example, identifies 18 variables as personally identifiable information. 45 C.F.R. § 164.514(b)(2)(i)(A)-(R). Disclosure experts have long understood that de-identification cannot guarantee anonymization, but this subtlety is lost in news reporting. For a discussion of reidentification risk and its treatment in the popular press, see Jane Yakowitz, *Tragedy of the Data Commons*, 25 Harv. J.L. & Tech. 1, 36-37 (2011).
- 19 Cf. Cynthia Dwork, A Firm Foundation for Private Data Analysis, 54 Commc'ns of the ACM 86, 89 (2011) (discussing the limited way the public uses interactive databases).
- 20 See Chin & Klinefelter, supra note 11, at 1452-53; Greengard, supra note 2, at 18; Ohm, supra note 10, at 1756-57; Wu, supra note 9, at 1137-38; Klarreich, supra note 7.
- 21 Cynthia Dwork, Presentation before the Eurostat Work Session on Statistical Data Confidentiality: Differentially Private Marginals Release with Mutual Consistency and Error Independent of Sample Size (Dec. 17-19, 2007), available at <http://www.unece.org/fileadmin/DAM/stats/documents/2007/12/confidentiality/wp.19.e.ppt>.
- 22 Id. (emphasizing this point on slide 24 of the accompanying PowerPoint presentation); see also Cynthia Dwork, Ask a Better Question, Get a Better Answer: A New Approach to Private Data Analysis, in Database Theory - ICDT 2007: 11th International Conference 18, 18-20 (Thomas Schwentick & Dan Suciu eds., 2006) (describing the Dinur-Nissim “blatant non-privacy” vulnerabilities and proposing differential privacy as a solution).
- 23 Irit Dinur & Kobbi Nissim, Revealing Information While Preserving Privacy, in Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems 202, 204, 206 (2003).
- 24 To be precise, if the largest amount of noise added is  $E$ , and if  $E$  is less than the number of data subjects, Dinur and Nissim showed that an adversary who could make unlimited numbers of queries could reconstruct a database so that the new database differed from the old database in no more than  $4E$  places. Thus, whenever  $E < n/400$ , the adversary will be able to construct a database that is accurate in 99% of the values, satisfying “blatant non-privacy.” Id. at 205-07.
- 25 Id. at 204.
- 26 Conditioned on the fact that  $E$  is no larger than  $\#n$ . Id. at 206.
- 27 For example,  $E = 50$  would avoid blatant non-privacy for a small database with 1000 subjects because the reconstructed database would be off in  $4 \times 50 = 200$  positions, rendering the database correct in only 80% of the values.
- 28 See John D. Cook, There Isn't a Googol of Anything, Endeavour (Oct. 13, 2010), <http://www.johndcook.com/blog/2010/10/13/googol/>; Googol, Wolfram Math World, <http://mathworld.wolfram.com/Googol.html> (last visited Jan. 29, 2014) (discussing the size of a google).
- 29 For example, theoretically nothing prevents a researcher from querying “what is the HIV status of subject #2502?” See Klarreich, supra note 7 (noting that differentially private data release algorithms allow adversaries to ask “practically any question about a database,” but “blur[s]” private information with noise).
- 30 Since a realistic adversary who is “bounded” or constrained by his computational ability will be thwarted by noise that is greater than  $\#n$ , large databases require comparatively less noise to overcome the reconstruction attack. For example, a database with 100 subjects

would require noise up to  $\pm 10$  to avoid such an attack (10% of the total number of subjects), but a database with 1,000,000 requires noise only up to  $\pm 1000$  (a tenth of one percent of the total number of subjects). See Dinur & Nissim, *supra* note 23, at 206.

- 31 Since the response to all queries are provided from the modified database, the best the adversary can hope to do is to reconstruct the modified database but not the original database.
- 32 A 99% accurate reconstruction is much more impressive when the binary outcomes are approximately equally likely (each outcome has probability approximately 50%). See Cynthia Dwork, *The Analytic Framework for Data: A Cryptographic View*, Microsoft Research 5 (2013), available at <http://cusp.nyu.edu/wp-content/uploads/2013/06/chapter11v2.pdf>.
- 33 See Tore Dalenius, *Towards a Methodology for Statistical Disclosure Control*, 5 *Statistisk tidskrift* 429, 432-33 (1977) (explaining that the context of the data refers to “[t]he frame: {O}E;” “[t]he data associated with the objects in the frame: I; C; X; Y, ...,Z;” “[t]he statistics released from the survey: S;” and “[t]he extra-objective data: E” and noting that “[i]f the release of the statistics S makes it possible to determine the value DK more accurately than is possible without access to S, a disclosure has taken place”).
- 34 See Krishnamurty Muralidhar & Rathindra Sarathy, *Security of Random Data Perturbation Methods*, 24 *ACM Transactions on Database Sys.* 487, 488 (1999); Rathindra Sarathy & Krishnamurty Muralidhar, *The Security of Confidential Numerical Data in Databases*, 13 *Info. Sys. Res.* 389, 393 (2002).
- 35 See, e.g., Lawrence H. Cox & John A. George, *Controlled Rounding for Tables with Subtotals*, 20 *Annals Operations Res.* 141, 141 (1989); Dalenius, *supra* note 33, at 441.
- 36 See, e.g., Muralidhar et al., *supra* note 15, at 1399; Muralidhar & Sarathy, *Data Shuffling*, *supra* note 15, at 658; D.B. Rubin, *Discussion of Statistical Disclosure Limitation*, 9 *J. Official Stat.* 461, 461 (1993).
- 37 Cynthia Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, in *Privacy, Security, and Trust in KDD-PinKDD 2007*, at 1, 1 (F. Bonchi et al. eds., 2008).
- 38 See *id.* at 5-6; Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 91; Cynthia Dwork, *Differential Privacy*, in 2 *Proceedings of the 33rd International Conference on Automata, Languages and Programming* 1, 8-9 (Michele Bugliesi et al. eds., 2006).
- 39 Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *supra* note 37, at 6; Dwork, *Differential Privacy*, *supra* note 38, at 9.
- 40 See Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *supra* note 37, at 6; Dwork, *Differential Privacy*, *supra* note 38, at 4.
- 41 See Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 89-90.
- 42 *Id.*
- 43 *Id.* at 89.
- 44 *Id.*
- 45 See *id.* at 87.
- 46 See *id.*
- 47 In 2010, the true figure was around 226,000. John Cook, *Millionaires to Double in Washington, but Will that Spark Angel Investment?*, *GeekWire* (May 4, 2011, 2:04 PM), <http://www.geekwire.com/2011/number-millionaires-double-washington-spark-angel-investment>.
- 48 For instance, the data producer may have added noise by selecting from random integer values in the range  $\pm 10$ . Hence, if the response to the query is 226,401, the adversary knows that Bill Gates is not a millionaire; if the response to the query is 226,422, the adversary knows that Bill Gates is a millionaire.

- 49 Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 90; Rathindra Sarathy & Krishnamurty Muralidhar, Some Additional Insights on Applying Differential Privacy for Numeric Data, in *Lecture Notes in Computer Science: Privacy in Statistical Databases 210, 211* (Josep Domingo-Ferrer & Emmanouil Magkos eds., 2011) [hereinafter Sarathy & Muralidhar, Additional Insights on Applying Differential Privacy].
- 50 The probability density function of a Laplace random variable is << equation >>.
- 51 In order to be able to compute  $\Delta f$ , a necessary step when implementing differential privacy, the data must have strict upper and lower bounds. Rathindra Sarathy & Krishnamurty Muralidhar, Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data, 4 *Transactions on Data Privacy* 1, 4 (2011) [hereinafter Sarathy & Muralidhar, Evaluating Laplace Noise]; Sarathy & Muralidhar, Additional Insights on Applying Differential Privacy, *supra* note 49; Larry Wasserman & Shuheng Zhou, A Statistical Framework for Differential Privacy, 105 *J. Am. Stat. Ass'n* 375, 378-79 (2010) (noting that ‘it is difficult to extend differential privacy to unbounded domains’’).
- 52 Dwork & Smith, *supra* note 14, at 140.
- 53 “[O]ur expected error magnitude is constant, independent of  $n$ .” Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 92.
- 54 Surely you remember from precalculus class that  $e^{\ln(2)} = 2$ , right?
- 55 See *supra* notes 42-43 and accompanying text.
- 56 The definition of differential privacy “trivially rules out the subsample-and-release paradigm discussed: For an individual  $x$  not in the dataset, the probability that  $x$ 's data is sampled and released is obviously zero; the multiplicative nature of the guarantee ensures that the same is true for an individual whose data is in the dataset.” Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91. Thus, the very release of microdata violates DP requirements. In addition, the application of differential privacy is a function of the query submitted, and since microdata is released so that a person may use it to issue any and all queries, the promises of differential privacy cannot be kept. Sarathy & Muralidhar, Evaluating Laplace Noise, *supra* note 51, at 3. To meet the differential privacy standard, even if it were possible, the data producer would have to add so much noise that the database would be meaningless. Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 92.
- 57 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91-92.
- 58 As we will demonstrate later in this Article, a data curator who wants to preserve even a small amount of data utility will have to choose a fairly large <<unknown symbol>>, allowing a generous tolerance for disclosure. See discussion *infra* Part III.D.
- 59 To understand why this is so, let's revisit the Bill Gates example. The adversary knows that 226,411 individuals have more than a million dollars in personal wealth. Issuing the query “How many individuals in Washington State have more than a million dollars?” may result in a response that has twice the probability that the true answer is 226,411 compared to the probability that the true answer is 226,412. The adversary can also issue the additional query “How many millionaires live in the 98039 zip code?,” which happens to be Bill Gates's zip code. Jeanne Lang Jones, The Sound's Wealthiest Zip Codes, *Puget Sound Bus. J.* (Feb. 6, 2005, 9:00 PM), <http://www.bizjournals.com/seattle/stories/2005/02/07/focus1.html>. Since the adversary has information on all millionaires in Washington State, we have to assume that he also knows all the million dollar income earners (other than Bill Gates) who live in this zip code. The response to this query may result in a response that, as in the previous query, suggests that the probability that Bill Gates is a millionaire is twice as likely as Bill Gates not a millionaire. Since the Laplace noise has been added independently, taken together, these two results provide the adversary with the assurance that the probability Bill Gates is a millionaire is four times as likely as the probability that he is not a millionaire. The privacy specification for the two queries combined is thus  $\ln(4) = 2 \times \ln(2) = 2 \times \text{unknown symbol}$  (twice the original <<unknown symbol>> we had set). In general, if the adversary is allowed to issue  $m$  queries and the privacy assurance is set to <<unknown symbol>> for each query, then for all  $m$  queries combined, the privacy assurance is only  $m \times \text{unknown symbol}$  (remember that a small <<unknown symbol>> provides more privacy). If we wish to limit the disclosure level to <<unknown symbol>> for all  $m$  queries combined, it would be necessary to set the disclosure level for each query to be (<<unknown symbol>>/  $m$ ). Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 92.
- 60 See, e.g., Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 89.

- 61 See *infra* Part III.E.
- 62 Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91.
- 63 See, e.g., Wu, *supra* note 9, at 1168-69.
- 64 Justin Brickell & Vitaly Shmatikov, The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing, in KDD '08 Proceedings of the 14th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining 70, 71 (2008) (“Sensitive attribute disclosure occurs when the adversary learns information about an individual's sensitive attribute(s). This form of privacy breach is different and in-comparable to learning whether an individual is included in the database, which is the focus of differential privacy.”); see also Wu, *supra* note 9, at 1121-23 (further clarifying the difference between research-based and data-based disclosures).
- 65 See Drew Olanoff, Zuckerberg on Building a Search Engine: Facebook Is Pretty Uniquely Positioned, at Some Point We'll Do It, TechCrunch (Sept. 11, 2012), <http://techcrunch.com/2012/09/11/zuckerberg-we-have-a-team-working-on-search> (stating that Facebook, for example, does over a billion queries a day).
- 66 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 92 (“Given any query sequence  $\langle\langle\text{unknown symbol}\rangle\rangle$ ,  $\langle\langle\text{unknown symbol}\rangle\rangle$ -differential privacy can be achieved by running  $K$  with noise distribution Lap  $\langle\langle\text{equation}\rangle\rangle$  on each query, even if the queries are chosen adaptively, with each successive query depending on the answers to the previous queries.”).
- 67 See *infra* Part III.D for a discussion of the multiple queries problem.
- 68 Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 90. The example has been repeated in other works, sometimes using Terry Gross instead of Alan Turing. See, e.g., Dwork & Smith, *supra* note 14, at 136.
- 69 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 90-91.
- 70 See Wu, *supra* note 9, at 1137-38. Consider the following example. Suppose Turing declares: “My salary is ten times the zip code of the White House.” Would publication of the White House's address violate Turing's privacy?
- 71 See Wu, *supra* note 9, at 1143-44. Felix Wu analogizes to the notions of cause-in-fact versus proximate cause. Disclosure of the external fact is a cause, but it is not a cause-in-fact. *Id.* at 1137-38.
- 72 Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91.
- 73 See Official Statistics Portal, Stat. Lith. (Apr. 9, 2014), <http://osp.stat.gov.lt/en/temines-lenteles19>.
- 74 See Average Female Height by Country, AverageHeight.co, <http://www.averageheight.co/average-female-height-by-country> (last visited Feb. 5, 2014).
- 75 See *infra* Part III.D for a discussion of the queries problem.
- 76 Selected Economic Characteristics: 2007-2011 American Community Survey 5-Year Estimates, US Census Bureau, [http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS\\_11\\_5YR\\_DP03](http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_5YR_DP03) (last visited Jan. 26, 2014).
- 77 Louise Story, Top Hedge Fund Managers Do Well in a Down Year, N.Y. Times, Mar. 24, 2009, <http://www.nytimes.com/2009/03/25/business/25hedge.html>.
- 78 See Selected Economic Characteristics: Booneville City, Kentucky, 2007-2011 American Community Survey 5-Year Estimates, US Census Bureau, <http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml> (search for “American Community Survey” and “Booneville city, Arkansas”; then show results from 2011) (last visited Feb. 5, 2014).
- 79 See *id.*
- 80 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91.
- 81 The fact that everyone knows with practical certainty that no one in the subset earned \$1 billion is irrelevant; the response distribution should be constructed in such a manner that \$1 billion income is feasible in this subset. See Dwork & Smith, *supra* note 14, at 137.



- 82 Note that this selection is less differential privacy-protecting, and thus more utility-preserving, than our last example.
- 83 See Selected Economic Characteristics: Booneville City, Kentucky, *supra* note 78.
- 84 For the purposes of this illustration, we have added noise only to the income variable. Adding noise to the number of residents would have made matters worse.
- 85 See Selected Economic Characteristics: 2007-2011 American Community Survey 5-Year Estimates, *supra* note 76.
- 86 We know that hedge fund operators like George Soros regularly take pay in excess of \$1 billion, so our illustration is a conservative estimate of the noise that would be added by differential privacy processes.
- 87 See Dwork, An Ad Omnia Approach to Defining and Achieving Private Data Analysis, *supra* note 37, at 7.
- 88 See *id.* at 8.
- 89 “Thus, our expected error magnitude is constant, independent of  $n$ .” Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 92.
- 90 See *id.*
- 91 See Selected Economic Characteristics: Booneville City, Kentucky, *supra* note 78.
- 92 *Id.*
- 93 One option for skewed data is to set arbitrary upper and lower limits for the values. For the income variable, it might be suggested that the upper limit should be set at (say) \$100 thousand. For this particular query, such a truncation would eliminate the problem of very large values. But the truncation would frustrate research on high income earners, or on income inequality. For example, if the query asked for the average income of hedge fund managers, truncating the upper limit of income at \$100 thousand would put nearly the entire data set in the truncated range. See J.K. Ord et al., Truncated Distributions and Measures of Income Inequality, 45 *Indian J. Stat.* 413, 414-15 (1983).
- 94 See Microsoft, *supra* note 4, at 5.
- 95 Assuming that the researcher sets the income in the middle of the range for each category, so that the 23 people earning between \$0 and \$10,000 are estimated to earn \$5,000, the 85 people earning between \$10,000 and \$50,000 are estimated to earn \$30,000, etc. The 5 people earning in excess of \$1 billion are estimated to earn \$1 billion and \$1. By this method, the researcher would reach an estimated average income over \$44 million. Using the same message using the “True Count” column would yield a more modest average income of \$35,254. We know that this is still quite far from the \$21,907 average that the Census reports for the town. See Selected Economic Characteristics: Booneville City, Kentucky, *supra* note 78.
- 96 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91.
- 97 It also seems to contradict the work of Dinur and Nissim, who conclude that in order to prevent blatant non-privacy, the noise added would have to be in the order of  $\epsilon$ . Dinur & Nissim, *supra* note 23, at 206.
- 98 Sandy Baum & Jennifer Ma, Education Pays: The Benefits of Higher Education for Individuals and Society, College Board Research Paper 10 (2007).
- 99 See Part I.B (discussing the need for the data curator to mask the presence or absence of any entry).
- 100 See, e.g., Chin & Klinefelter, *supra* note 11, at 1452-53; Ohm, *supra* note 10, at 1756; Wu, *supra* note 9, at 1139-40; Felten, *supra* note 12.
- 101 Cynthia Dwork et al., Differentially Private Marginals Release with Mutual Consistency and Error Independent of Sample Size, Eurostat Work Session on Stat. Data Confidentiality 193, 198 (2007).
- 102 See Greengard, *supra* note 2, at 17; Chin & Klinefelter, *supra* note 11, at 1452-55.

- 103 See Chin & Klinefelter, *supra* note 11, at 1422-23. Chin and Klinefelter describe an investigation that they conducted to assess the security practices of Facebook. *Id.* at 1432-45. Based on their analysis, the authors conclude that Facebook is likely using differential privacy, even though Facebook has never indicated that they are. *Id.* at 1422-23. Since the researchers submitted over 30,000 queries, almost any selection of epsilon would have required the noise for each query to dominate the true answer. See *id.* at 1436. Either Facebook is using some other noise-adding mechanism, or the company is implementing differential privacy incorrectly.
- 104 Klarreich, *supra* note 7.
- 105 See *supra* Part II.A.
- 106 See Klarreich, *supra* note 7.
- 107 See Microsoft, *supra* note 4, at 4-5.
- 108 *Id.* at 5.
- 109 Remember that, because the effect on privacy of queries is cumulative, the noise added to each successive query must increase in order to satisfy differential privacy for any specific overall selection of  $\epsilon$ . See *supra* note 59 and accompanying text.
- 110 The noise will be randomly selected from the distribution generated by the Laplace function  $Lap(\epsilon) = Lap(910.239)$ .
- 111 Astute readers may notice that the random realization reported in Table 10 is very similar to the output that our fictional internist was confronting in the Introduction. See *supra* note 3 and accompanying text. Indeed, we took the same error drawn here and added it to our equally fictional “true” responses, which was 20 for each year. Thus, as it turns out, this internist would have had little to worry about if she had known the truth--that seeing a few cases over the course of several weeks is par for the course. Since the internist did not know the true values, though, she would have had little reason to feel comforted or alarmed by the responses that she received.
- 112 There are also some situations in which restricting the database to a small number of queries in order to reduce the magnitude of the noise can produce disclosures. For an example, see Cormode, *supra* note 2, at 1254. These disclosures are not, technically, within Dwork’s definition of “disclosure” motivating her differential privacy solutions.
- 113 Klarreich, *supra* note 7.
- 114 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91.
- 115 See *id.*
- 116 See *id.*
- 117 See Klarreich, *supra* note 7; Chin & Klinefelter, *supra* note 11, at 1433-35.
- 118 See, e.g., Wu, *supra* note 9, at 1138.
- 119 *Id.*
- 120 See *supra* Part II.D.
- 121 Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 92.
- 122 See Microsoft, *supra* note 4, at 5.
- 123 *Id.*
- 124 Kobi Nissim et al., Smooth Sensitivity and Sampling in Private Data Analysis, in STOC’07 Proceedings of the 39th Annual ACM Symposium on Theory of Computing 75, 78 (David S. Johnson & Uriel Feige eds., 2007).
- 125 See Bhaskar et al., *supra* note 2, at 216 (“The amount of noise introduced in the [differentially private] query-response is ... [i]ndependent of the actual data entries ...”).

- 126 Felten, *supra* note 12.
- 127 Recall that the differentially private noise is independent from the size of the database so that the reported answer approaches the true answer as the size increases.
- 128 “The average US panel size is about 2,300.” Justin Altschuler, MD, David Margolius, MD, Thomas Bodenheimer, MD & Kevin Grumbach, MD, Estimating a Reasonable Patient Panel Size for Primary Care Physicians with Team-Based Task Delegation, 10 *Annals Fam. Med.* 396, 396 (2012).
- 129 The 1% to 99% range of the noise would be approximately 40,000 to +40,000.
- 130 See Wu, *supra* note 9, at 1137-40; Felten, *supra* note 12.
- 131 Tore Dalenius, Towards a Methodology for Statistical Disclosure Control, 5 *Statistisk tidskrift* 429, 433 (1977).
- 132 *Id.* at 439-40 (“It may be argued that elimination of disclosure is possible only by elimination of statistics.”).
- 133 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91-92.
- 134 See *id.* at 88-89.
- 135 Dwork, An Ad Omnia Approach to Defining and Achieving Private Data Analysis, *supra* note 37, at 7.
- 136 Daniel Kifer & Ashwin Machanavajjhala, No Free Lunch in Data Privacy, in *SIGMOD '11 Proceedings of 2011 ACM SIGMOD International Conference on Management of Data* 193, 193 (2011).
- 137 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 92.
- 138 See Olanoff, *supra* note 65.
- 139 See Dwork & Smith, *supra* note 14, at 137; Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91.
- 140 See Dwork & Smith, *supra* note 14, at 137; Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 92. Returning to Chin and Klinefelter's analysis of responses to 30,000+ different Facebook queries, Chin and Klinefelter conclude that Facebook is likely using a rounding function and a noise addition mechanism that is consistent with  $\epsilon = 0.181$  for each query. Chin & Klinefelter, *supra* note 11, at 1433-40. For the set of 30,000+ queries as a whole, this would imply that  $\epsilon = (0.181 \times 30000) = 5430$  which translates into a privacy risk ratio of  $e^{5430}$  which is so large that, for all practical purposes, it might as well be infinite. Whether the mistake is Chin and Klinefelter's (for misidentifying differential privacy) or Facebook's (for misapplying it), it shows a frequent, critical failure to understand that the response to every query contributes to the adversary's ability to compromise the privacy of an individual, resulting in wildly overstated descriptions of the privacy offered by differential privacy mechanisms.
- 141 See Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 88.
- 142 Even seasoned researchers make the mistake of setting unreasonably high values for  $\epsilon$ . For instance, Anne-Sophie Charest sets  $\epsilon = 250$ , and David McClure and Jerome Reiter set  $\epsilon = 1000$ , which offers no guarantee of privacy whatsoever. See Anne-Sophie Charest, How Can We Analyze Differentially-Private Synthetic Datasets?, 2 *J. Privacy & Confidentiality* 21, 27 (2010); David McClure & Jerome P. Reiter, Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data, 5 *Transactions on Data Privacy* 535, 536 (2012).
- 143 One of the interesting aspects of the Laplace distribution is that the noise for  $m$  queries is a direct multiple of the noise for one query. The Laplace inverse cumulative distribution function with mean zero is written as:  $\frac{1}{2b} \exp(-|x|/b)$  where  $b$  is the shape parameter of the Laplace distribution and  $p$  is a random number between 0 and 1. When a single query is answered,  $\frac{1}{2b} \exp(-|x|/b)$  and when  $m$  queries are answered  $\frac{1}{2b} \exp(-|x|/b)$ . For a given random number  $p$ , the noise using  $b'$  is  $m$  times the noise generated using  $b$ .
- 144 Dwork, A Firm Foundation for Private Data Analysis, *supra* note 19, at 91 (emphasis omitted).
- 145 See *id.* at 91-92.

- 146 Graham Cormode also provides an interesting example of a disclosure that can be made while satisfying differential privacy, but which is avoidable with more traditional, context-driven privacy measures. Cormode, *supra* note 2, at 1256-57.
- 147 See Dwork, *A Firm Foundation for Private Data Analysis*, *supra* note 19, at 92.
- 148 Even tabular data has the potential to cause confusion. Klarreich, author of the *Scientific American* article, provides an illustration of a type of disclosure that occurs with genotype frequencies. Klarreich, *supra* note 7. Unfortunately, in this situation, it would not be possible to maintain the privacy parameter for each cell and the overall database at  $\epsilon$ . The data involves frequencies of thousands of different single nucleotide polymorphisms (SNPs) and every individual is represented in every SNP frequency. See *id.* The addition/deletion of one record will modify every one of the SNP frequencies. To see an attack taking advantage of these circumstances, see Daniel I. Jacobs et al., *Leveraging Ethnic Group Incidence Variation to Investigate Genetic Susceptibility to Glioma: A Novel Candidate SNP Approach*, 3 *Frontiers in Genetics* 203, 203 (2012).
- 149 Hopefully the curator's source for learning the global range does not employ differential privacy.
- 150 See Dwork & Smith, *supra* note 14, at 139. Mathematically, the relationship looks like this:  $\frac{1}{\epsilon} \leq \frac{P(S)}{P(S')}$  where  $\delta$  is small.
- 151 The extent to which actual probability ratio is different from the ratio that includes or excludes a data subject is bounded by the  $\frac{1}{\epsilon}$ , but when  $\frac{1}{\epsilon}$  is very small (say 0.00001) and  $\delta = 0.01$ , the privacy ratio can exceed differential privacy standards by 1000. Even though  $\delta$  is small, the risk of disclosure can be very large.
- 152 Ashwin Machanavajjhala et al., *Privacy: Theory Meets Practice on the Map*, in *ICDE '08 Proceedings of the 2008 IEEE 24th International Conference on Data Engineering* 277, 283 (2008).
- 153 See *id.* at 280-81.
- 154 For example, the authors go on to propose a relaxation of differential privacy that satisfies differential privacy albeit with  $\epsilon = 8.6$ , which implies a privacy risk ratio of  $e^{8.6} = 5431.66$ . *Id.* at 284. This implies that, based on the responses (or in this case released data), we can conclude the presence of an individual has probability that can be 5431.66 times higher than the absence of an individual.
- 155 See Dwork, *An Ad Omnia Approach to Defining and Achieving Private Data Analysis*, *supra* note 37, at 1 (criticizing disclosure prevention mechanisms for being syntactic and ad hoc).
- 156 See Machanavajjhala et al., *supra* note 152, at 277.
- 157 See, e.g., Bhaskar et al., *supra* note 2, at 216 (“While the form of our guarantee is similar to DP, where the privacy comes from is very different, and is based on: 1) A statistical (generative) model assumption for the database, 2) Restrictions on the kinds of auxiliary information available to the adversary.”).
- 158 For example, differential privacy offers no greater security against Dinur-Nissim “blatant non-privacy” unless the data curator strictly limits the number of queries that can be issued to the system. Cf. Dinur & Nissim, *supra* note 23, at 203-04, 206. Other noise-adding approaches, too, can avoid the Dinur-Nissim results by limiting the number of queries. See *supra* note 28 and accompanying text.
- 159 See Barbara J. Evans, *Much Ado About Data Ownership*, 25 *Harv. J.L. & Tech.* 69, 76, 94 (2011) (discussing the value of compiling patient metadata for research).

## Differential Privacy and Census Data: Implications for Social and Economic Research<sup>†</sup>

By STEVEN RUGGLES, CATHERINE FITCH, DIANA MAGNUSON,  
AND JONATHAN SCHROEDER\*

In September 2018, the Census Bureau announced a new set of methods for disclosure control in public use data products, including aggregate-level tabular data and microdata derived from the decennial census and the American Community Survey (ACS) (US Census Bureau 2018a). The new approach, known as differential privacy, “marks a sea change for the way that official statistics are produced and published” (Garfinkel, Abowd, and Powazek 2018, p. 136).

In accordance with census law, for the past six decades the Census Bureau has ensured that no census publications allow specific census responses to be linked to specific people. Differential privacy requires protections that go well beyond this standard; under the new approach, responses of individuals cannot be divulged even if the identity of those individuals is unknown and cannot be determined. In its pure form, differential privacy techniques could make the release of scientifically useful

microdata impossible and severely limit the utility of tabular small-area data.

Initially, the Census Bureau plans to apply differential privacy techniques to the two most intensively-used sources in social science and policy research, the ACS and the decennial census (US Census Bureau 2018b). These data generate some 17,000 publications each year. The ACS and decennial census are widely used in analyses of the economy, population change, and public health, and they are indispensable tools for federal, state and local planning. Common topics of analysis include poverty, inequality, immigration, internal migration, ethnicity, residential segregation, transportation, fertility, nuptiality, occupational structure, education, and family change. The data are routinely used to construct contextual measures that control for neighborhood effects on health and disease. Investigators exploit policy discontinuities across time and space, disasters, and weather events as natural experiments that allow causal inferences. Policymakers and planners use small-area data from the ACS and decennial census to understand local environments and focus resources where they are needed. Businesses use the data to estimate future demand and determine business locations.

Adoption of differential privacy will have far-reaching consequences for users of the ACS and decennial census. It is possible—even likely—that scientists, planners, and the public will soon lose the free access we have enjoyed for the past six decades to reliable public Census Bureau data describing American social and economic change.

The differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau (Ruggles et al. 2018). By imposing unrealistic disclosure rules, the Census Bureau may be forced to lock up data that are indispensable for basic

\* Ruggles: IPUMS, University of Minnesota, 50 Willey Hall, 225 19th Avenue S., Minneapolis, MN 55455 (email: [ruggles@umn.edu](mailto:ruggles@umn.edu)); Fitch: IPUMS, University of Minnesota, 50 Willey Hall, 225 19th Avenue S., Minneapolis, MN 55455 (email: [fitch@umn.edu](mailto:fitch@umn.edu)); Magnuson: Bethel University, St. Paul, MN 55112 (email: [d-magnuson@bethel.edu](mailto:d-magnuson@bethel.edu)); Schroeder: IPUMS, University of Minnesota, 50 Willey Hall, 225 19th Avenue S., Minneapolis, MN 55455 (email: [jps@umn.edu](mailto:jps@umn.edu)). Support for this work was provided by the Minnesota Population Center at the University of Minnesota (P2C HD041023). We are grateful for the comments and suggestions of Margo J. Anderson, J. Trent Alexander, Wendy Baldwin, Jane Bambauer, John Casterline, Sara Curran, Michael Davern, Roald Euler, Reynolds Farley, Katie Genadek, Miriam L. King, Wendy Manning, Douglas Massey, Robert McCaa, Frank McSherry, Krish Muralidhar, Samuel Preston, Matthew Sobek, Stewart Tolnay, David Van Riper, and John Robert Warren.

<sup>†</sup>Go to <https://doi.org/10.1257/pandp.20191107> to visit the article page for additional materials and author disclosure statement(s).

research and policy analysis. If public use data become unusable or inaccessible because of overzealous disclosure control, there will be a precipitous decline in the quantity and quality of evidence-based policy research.

### I. Differential Privacy and Census Law

Differential privacy guarantees that the presence or absence of any individual case from a database will not significantly affect any database query. In particular, “even if the participant removed her data from the dataset, no outputs ... would become significantly more or less likely” (Dwork 2006, p. 9). This definition has the advantage of being relatively simple to formalize, and that formalization yields a metric summarizing a database’s level of “privacy” in a single number ( $\epsilon$ ).

The application of differential privacy to census data represents a radical departure from established Census Bureau confidentiality laws and precedents (Ruggles et al. 2018). The differential privacy requirement that database outputs do not significantly change when any individual’s data is added or removed has profound implications. In particular, under differential privacy it is prohibited to reveal characteristics of an individual even if the identity of that individual is effectively concealed.

As the Census Bureau acknowledges, masking respondent characteristics is not required under census law. Instead, the laws require that the identity of particular respondents shall not be disclosed. In 2002, Congress explicitly defined the concept of identifiable data: it is prohibited to publish “any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.”<sup>1</sup>

For the past six decades the Census Bureau disclosure control strategy has focused on targeted strategies to prevent re-identification attacks, so that an outside adversary cannot positively identify which person provided a particular response. The protections in place—sampling, swapping, suppression of geographic information and extreme values, imputation, and perturbation—have worked extremely well to meet this standard. Indeed, there is not a single

documented case of anyone outside the Census Bureau revealing the responses of a particular identified person in public use decennial census or ACS data.

### II. Reconstruction and Re-identification

Census analysts argue that new disclosure rules are needed because of the threat of “database reconstruction.” Database reconstruction is a process for inferring individual-level responses from tabular data. Abowd (2017, p. 10) argues that database reconstruction “is the death knell for public-use detailed tabulations and microdatasets as they have been traditionally prepared.”

The Census Bureau conducted a database reconstruction experiment that sought to identify the age, sex, race, and Hispanic origin for the population of each of the 6.3 million inhabited census blocks in the 2010 census. According to Abowd (2018a, p. 6), the experiment confirmed “that the micro-data from the confidential 2010 Hundred-percent Detail File (HDF) can be accurately reconstructed” using only the public use summary tabulations. The HDF is the individual-level complete census incorporating confidentiality protections such as swapping similar households that reside in different places.

It should not be a great surprise that individual-level characteristics can be inferred from tabular data. Any table that includes data about people can be rearranged as individual-level data. For the Census Bureau database reconstruction experiment, analysts started with a table of age by sex by race by Hispanic origin, and converted the table to microdata. For example, if a particular census tract had three black non-Hispanic women aged 25 to 29, they created three microdata records with these individual-level characteristics. By repeating this process for every cell in the table, the full content of the table may be expressed in the form of microdata. Then the Census Bureau added more detail on place of residence, age, and race by cross-referencing across multiple tables.

The reliability of the method varies depending on the characteristics of the census block. For some blocks, there are multiple possible solutions, making inferences difficult (Abowd 2018b). In other cases it is easy to infer

<sup>1</sup> Title 5 USC, §502 (4), Public Law 107–347.

individual-level variables. For example, 47 percent of blocks contain a single race and 60 percent have a single Hispanic (or non-Hispanic) ethnicity; accurately inferring race or ethnicity for persons in such homogeneous blocks is trivial. Once the individual-level data were fully reconstructed, the Census Bureau tested the accuracy by matching the reconstructed individual-level records to the microdata that had been used to create the public use tables. For each individual in the reconstructed dataset, the software searched the original microdata for a person with a matching age, sex, race, and Hispanic origin.

In the end, only 50 percent of the reconstructed cases accurately matched a case from the HDF source data (Abowd 2018c; Hansen 2018). In the great majority of the mismatched cases, the errors resulted from a discrepancy in age. Given the 50 percent error rate, it is not justifiable to describe the microdata as “accurately reconstructed” (Abowd 2018a, p. 6).

Reconstructing microdata from tabular data does not by itself allow identification of respondents; to determine who the individuals actually are, one would then have to match their characteristics to an external identified database (including, for example, names or Social Security numbers) in a conventional re-identification attack. The Census Bureau attempted to do this but only a small fraction of re-identifications actually turned out to be correct, and Abowd (2018d, p. 15) concluded that “the risk of re-identification is small.” Therefore, the system worked as designed: because of the combination of swapping, imputation and editing, reporting error in the census, error in the identified credit agency file, and errors introduced in the microdata reconstruction, there is sufficient uncertainty in the data to make positive identification by an outsider impossible.

### III. Implications for Tabular Data

Despite the low risk of re-identification in the Census Bureau experiment, the 100 percent tabular data from the decennial census pose some special disclosure control challenges. Because these tables include the entire population with very fine geographic detail, there could be potential for re-identification if no disclosure protections were applied.

The block-level decennial tables include very few variables, and the research applications of these tables are comparatively limited. The Census Bureau has not yet demonstrated that differential privacy is the most effective and efficient means of preventing positive re-identification while maximizing utility of these data. It is nevertheless possible that some variant of differential privacy or a similar method could be applied that would preserve usability for the relatively limited applications of the block data while strengthening disclosure control.

Differentially-private tabular data from the ACS is considerably more challenging than the 100 percent files, because there are many more variables and the data are used for a much wider range of research and planning purposes. It may be impossible to create a differentially-private version of the ACS tables that would meet the needs of researchers and planners. Fortunately, tabular data from the ACS have features that make them inherently less identifiable than the 100 percent census data. The ACS is a sample with just 1.5 percent of the population each year, and there is no block-level data. At the block group level, the ACS data must combine five years of data, so there is temporal as well as spatial uncertainty. The chances of any particular respondent being included in the file are very low. If an exact match is found through a reconstruction and re-identification attack, it would be impossible to determine whether the match was correct because there may be another exact match which was not sampled. Accordingly, less aggressive disclosure controls may be appropriate for ACS tabular data.

### IV. Implications for Microdata

Differentially private microdata is not a realistic disclosure control solution. ACS microdata samples directly provide individual-level characteristics derived from real people, and this in itself represents a violation of the core principles of differential privacy (Bambauer, Krishnamurty, and Sarathy 2014). A recent paper published by Census Bureau privacy experts notes that “record-level data are exceedingly difficult to protect in a way that offers real privacy protection while leaving the data useful for unspecified analytical purposes. At present, the Census Bureau advises research users who

require such data to consider restricted-access modalities,” in particular the Federal Statistical Research Data Centers (Garfinkel, Abowd, and Powazek 2018, p. 138). By “real privacy protection,” the authors mean differential privacy, not confidentiality protection as defined in census law and precedent. By “unspecified analytical purposes” the authors mean any analytic purposes that are not anticipated in advance.

To guarantee differential privacy, microdata must be simulated using statistical models rather than directly derived from the responses of real people (Dajani et al. 2017, Reiter forthcoming). Such modeled data—usually called synthetic data—captures relationships between variables only if they have been intentionally included in the model. Accordingly, synthetic data are poorly suited to studying unanticipated relationships, which would greatly impede new discoveries from differentially private microdata.

Census Bureau privacy researchers argue that if the public use data become unusable, scientific research can be carried out in the secure Federal Statistical Research Data Centers (FSRDCs). This is not a practical plan. As we have argued elsewhere, the FSRDC network would have to be expanded by several orders of magnitude to accommodate the volume of research now carried out using public use microdata, and most projects would be ineligible (Ruggles et al. 2018). Without major legal changes and a massive infusion of funds, restricted access is not a viable alternative to public use microdata.

The existing ACS microdata samples provide powerful protections against re-identification. The public use microdata are a sample of a sample; annual information on less than 1 percent of the population is released to the public. There is no geographic identification of places with fewer than 100,000 inhabitants. Outlying values are top-coded or bottom-coded; variables are grouped into categories representing at least 10,000 persons in the general population; ages are perturbed for some population subgroups; and additional noise is added for persons in group quarters or with rare combinations of characteristics. These measures have proven highly effective. It is impossible for an intruder to determine whether any attempted re-identification was successful, or even to calculate the odds that the attempt was successful. Accordingly, we recommend only incremental

improvements in disclosure control for the ACS microdata samples.

## V. Discussion and Recommendations

There are compelling reasons to take confidentiality protection seriously. Re-identification is a greater concern today than in the past, both because of the declining cost of computing and the increasing availability of private-sector identified data that might be used in an attack. For the past two decades, the Census Bureau has conducted systematic evidence-based research on the actual risks of re-identification in public use census data (Ruggles et al. 2018). This empirical approach targets methods of disclosure control that address realistic threats by focusing on particular population subgroups and variables posing the greatest risks, while minimizing damage to data utility. The Census Bureau should build on this work by continuously modernizing and strengthening its disclosure control methods.

Differential privacy goes far beyond what is necessary to keep data safe under census law and precedent. Differential privacy focuses on concealing individual characteristics instead of respondent identities, making it a blunt and inefficient instrument for disclosure control. As Abowd and Schmutte (2019) have observed, there is a trade-off between privacy and data usability. As defined by census law, privacy means protecting the identity of respondents from disclosure. The core metric of differential privacy, however, does not measure risk of identity disclosure (McClure and Reiter 2012). Because differential privacy cannot assess disclosure risk as defined under census law and precedent, it cannot be used to optimize the privacy/usability trade-off.

The United States is facing existential challenges. We must develop policies and plans to adapt to accelerating climate change; that will require reliable ACS microdata and small area data. The impact of immigration—one of the most divisive issues in American policy debates—cannot be measured without the ACS tables and microdata. More broadly, investigators need data to investigate the causes and consequences of rapidly growing inequality in income and education. We need to examine how fault lines of race, ethnicity, and gender are dividing the country. We need basic data to study the shifts in spatial organization of the



population that are contributing to fragmentation of politics and society. This is not the time to impose arbitrary and burdensome new rules, with no basis in law or precedent, which will sharply restrict or eliminate access to the nation's core data sources.

The Census Bureau's mission is "to serve as the nation's leading provider of quality data about its people and economy" (US Census Bureau 2018c, p. 3). To meet that core responsibility, the Census Bureau must make accurate and reliable data available to the public. The Census Bureau has an extraordinary record—better than anywhere else in the world—of making powerful public use data broadly accessible. Just as important, the Census Bureau also has an unblemished record of protecting confidential information. There are no documented instances in which the identity of a respondent to the decennial census or ACS has been positively identified by anyone outside the Census Bureau using public use data. We must ensure that both of these powerful traditions continue. We need both broad democratic access to high-quality data and strong confidentiality protections to understand and overcome the daunting challenges facing our nation and the world.

We have three specific recommendations:

- (i) *Differential privacy might be feasible for summary files, but more testing is needed.* The most plausible use of the technique is for the 100 percent tabular files, where the range of applications is relatively limited. Making useful differentially private ACS tabular data will be challenging and may not be practical.
- (ii) *To preserve the utility of public use microdata, the Census Bureau should pursue alternative disclosure control strategies.* Differential privacy is not appropriate for ACS microdata. Differentially private synthetic microdata are not suitable for most original research problems. There is no legal mandate for differential privacy, and restricted-access alternatives to public use data are not feasible.
- (iii) *The Census Bureau should proceed cautiously in close consultation with the user community.* If new disclosure control technology is rushed out prematurely

and without adequate evaluation, damaging mistakes are inevitable. For any new disclosure control procedures, the research community should have an opportunity to test the methods through a rigorous process before they are finalized. The best way to achieve this is by enlisting the research community to replicate past peer-reviewed research using data that incorporate new disclosure control methods.

## REFERENCES

- Abowd, John M.** 2017. "Research Data Centers, Reproducible Science, and Confidentiality Protection: The Role of the 21<sup>st</sup> Century Statistical Agency." Paper presented at the Census Scientific Advisory Committee Meeting, Suitland, MD.
- Abowd, John M.** 2018a. "How Modern Disclosure Avoidance Methods Could Change the Way Statistical Agencies Operate." Paper presented at the Federal Economic Statistics Advisory Committee Meeting, Suitland, MD.
- Abowd, John M.** 2018b. "Staring-Down the Database Reconstruction Theorem." Paper presented at the Joint Statistical Meetings, Vancouver, BC.
- Abowd, John M.** 2018c. Personal communication, December 11, 2018.
- Abowd, John M.** 2018d. "The U.S. Census Bureau Adopts Differential Privacy." Paper presented at the 24<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London.
- Abowd, John M., and Ian M. Schmutte.** 2019. "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices." *American Economic Review* 109 (1): 171–202.
- Bambauer, Jane, Krishnamurthy Muralidhar, and Rathindra Sarathy.** 2014. "Fool's Gold: An Illustrated Critique of Differential Privacy." *Vanderbilt Journal of Entertainment and Technology Law* 16 (4): 701–55.
- Dajani, Aref N., Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, et al.** 2017. "The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau." Paper presented at the Census Scientific Advisory Committee Meeting, Suitland, MD.

- Dwork, Cynthia.** 2006. "Differential Privacy." In *Automata, Languages and Programming: 33<sup>rd</sup> International Colloquium*, edited by Michele Bugliesi, Bart Preneel, Vlarimiro Sassone, and Ingo Wegener, 1–12. Heidelberg: Springer.
- Garfinkel, Simson L., John M. Abowd, and Sarah Powazek.** 2018. "Issues Encountered Deploying Differential Privacy." In *2018 Workshop on Privacy in the Electronic Society*, edited by Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter, 133–37. New York: ACM.
- Hansen, Mark.** 2018. "To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data." *New York Times*, December 5. <https://www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html>.
- McClure, David, and Jerome P. Reiter.** 2012. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." *Transactions on Data Privacy* 5: 535–52.
- Reiter, Jerome P.** Forthcoming. "Differential Privacy and Federal Data Releases." *Annual Review of Statistics and Its Application*.
- Ruggles, Steven, et al.** 2018. "Implications of Differential Privacy for Census Bureau Data and Scientific Research." Minnesota Population Center Working Paper 2018-6.
- US Census Bureau.** 2018a. "Statistical Safeguards." Data Protection and Privacy Program. [https://www.census.gov/about/policies/privacy/statistical\\_safeguards.html](https://www.census.gov/about/policies/privacy/statistical_safeguards.html).
- US Census Bureau.** 2018b. "Restricted-Use Microdata." US Census Bureau, [https://www.census.gov/research/data/restricted\\_use\\_microdata.html#CRE1](https://www.census.gov/research/data/restricted_use_microdata.html#CRE1).
- US Census Bureau.** 2018c. *Strategic Plan—Fiscal Year 2018 through Fiscal Year 2022*. Suitland, MD: US Census Bureau.

# MEMORANDUM

To: The Honorable Ralph Northam  
Governor of the Commonwealth of Virginia

From: Meredith Strohm Gunter, Director, Strategy and Public Engagement  
Weldon Cooper Center for Public Service, University of Virginia

Date: January 23, 2020

Re: 2020 Census Data Distortion

---

While the importance of the 2020 Census is fully recognized, most census data users have not yet heard about “differential privacy,” a new mathematical procedure devised by the Census Bureau that will be applied to the 2020 Census data before it is released to enhance data privacy protection. Our analysis indicates that data accuracy at the sub-state (region, county, city, town) level will be sacrificed as a result of this new approach to data release. This inaccuracy may lead to misallocation of funds, poor capacity for planning, substandard service provision, and a competitive disadvantage in economic and workforce development.

For example, working with data provided by the Bureau to demonstrate the effects of their new procedure, we found the total number of girls ages 15-19 in the City of Emporia were decreased from the actual 185 to only 30. Applying this number to the teen pregnancy rate for Emporia increased the rate from 10 percent to 66 percent. This is not only ludicrous, but, if consistent across localities and subject areas, deeply damaging to the ability of state and local governments and non-profits to accurately address the needs of Virginians.

According to the Census Bureau’s current plan for the 2020 Census, an accurate headcount will only be available at the state level (in order to serve the fundamental purpose of congressional re-apportionment). The headcounts for counties, cities, and towns, as well as population characteristics, such as age, gender, race/ethnicity will be injected with data noise so that no individual information can be reconstructed.

As a result, none of the sub-state numbers would be actual counts, but rather a noise-injected proxy. The demonstration data (using the 2010 Census) provided a preview of the consequential changes. Shifts are almost always from large groups to small groups, and this pattern is not random. Since the state total must be held constant, population among localities is a zero-sum game, and the algorithms being tested shift population from urban to rural areas, and from large race groups to small race groups. A rural, declining, old, predominant white community, for example, may appear instead growing, younger, and more diverse. Distortion in age groups is the reason for the Emporia distortion mentioned above.

The data distortion has multiple concerning effects:

1. Redistricting data will be inaccurate, both in terms of the actual size of the voting age population in each census block and their racial characteristics. Majority-minority districts could lose their status due to noise injection, and the reverse could also come to pass.
2. Funding equity across localities will be severely impaired. While federal dollars to each state will be equitable because the state population will reflect the actual census count, money going to each community and program will not, as their population totals will be distorted. The targeted population of each funding program could artificially become smaller or larger, undermining program effectiveness and resources.
3. Many federal, state, and local statistics will produce inconsistent, unreasonable results, as they rely on the census count as a benchmark. Health, education, and criminal justice, for example, heavily rely on age-, gender-, race-specific census data to derive statistically sound rates that may be compared over time. The noise injection will make such rates incomprehensible and comparisons across geography and time meaningless.
4. Government services will be significantly impacted. Housing, transportation, emergency management, to name just a few, need accurate census data for planning, budgeting, and program delivery.

Data user communities across the country have voiced grave concerns about the Census Bureau's differential privacy procedure. It is detrimental to data accuracy, and to the status of the census data as the gold standard. The planned data distortion will last for the entire decade and carries implications that will be felt far and wide.

As a thought leader in the country, your steadfast support for a complete count of Virginia residents in the census has been inspiring. Full participation by Virginians will deliver high quality data to the Census Bureau. Now we need to make certain that data is reported out accurately.

We would be happy to assist an effort by your administration to bring greater awareness of this issue to governors and other state and local leaders through the National Governors' Association, the National League of Cities, and the National Association of Counties. We could also work with your administration to urge state and local leaders across Virginia to evaluate the proposed plans by the Bureau and to express their opinions through the email address and process identified on the enclosed.

Thank you for your leadership for our Commonwealth, and your attention to this issue.

## Resources relevant to the Census Bureau’s proposed differential privacy initiative

### VIRGINIA CONTACTS

Dr. Meredith Strohm Gunter  
Director, Strategy and Public Engagement, Weldon Cooper Center for Public Service  
University of Virginia  
[Meredith.gunter@virginia.edu](mailto:Meredith.gunter@virginia.edu) 434-982-5585

Dr. Qian Cai (pronounced “Chien Sigh”)  
Director, Weldon Cooper Center Demographics Research Group, University of Virginia  
Virginia’s state representative to the Census Bureau Federal-State Cooperative Program  
for Population Estimates  
[qian.cai@virginia.edu](mailto:qian.cai@virginia.edu) 434-982-5581

### RESOURCES

AP story: <https://federalnewsnetwork.com/big-data/2019/12/researchers-warn-census-about-accuracy-concerns-with-method-2/>

Census Bureau comment page: <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html>

Census Bureau email address for comments:  
[dcmd.2010.demonstration.data.products@census.gov](mailto:dcmd.2010.demonstration.data.products@census.gov)

Social Explorer uses cookies to allow us to better understand how the site is used and to improve your experience. By continuing to use this site, you consent to this policy. [About cookies](#)



## Controversial Census Bureau Plan That Makes Data Less Accurate Goes to Court

SUNDAY, MAY 02, 2021

[Tweet](#)

As some states celebrate and others reel from the release of the state population figures, a storm continues around the Census Bureau's unprecedented and massively controversial plan that makes the Census data inaccurate, purportedly to make it less vulnerable to privacy attacks. A hearing will take place Monday in an Alabama federal courtroom before a three-judge panel (with the possibility for a direct appeal to the U.S. Supreme Court). It could determine how useful the data will be when it is released and could therefore have an impact on the future of redistricting, federal funding, public policy decisions and more. ([Filings for the case are available here.](#))

Redistricters across the political spectrum, voting rights advocates, legislators, and legislative staff members, as well as mayors, town and village supervisors and other parties are very concerned about the accuracy of the soon-to-be-released Census data. Letters expressing these concerns have been sent to the Census Bureau by many state demographers, a wide variety of researchers, redistricting professionals, civil rights organizations, and others. Aside from the idea that the data may not be usable for its main intended purpose of redistricting, the Alabama case argues that the Census ignored the requirement to consult with the states and obtain agreement on the data to be released. This consultation never happened, and the Census Bureau is relying upon its own assessment to decide what is released. Disagreement over the new data plan exists



1

the Bureau as well — many internal Census documents submitted in the case show that  
<https://www.socialexplorer.com/blog/post/controversial-census-bureau-plan-that-makes-data-less-accurate-goes-to-court-11452>

the Bureau as well as many internal Census documents submitted in the case show that

Census staff were also quite concerned about this new approach to “protecting” census data.

The Census response, so far, has been a series of “demonstration products” based upon the 2010 Census, which researchers must assess to see if the latest version of these is “fit for use” as the Census describes it. The [most recent set](#) of these was released on April 28, 2021.

The first wave of detailed Census data is scheduled to be released in mid- to late-August. (This release is colloquially called the PL 94-171 files, named after the 1975 statute that set up the standards for constructing the files after each census.) The main official use for those data is to draw legislative districts from the congressional level down to the city and village council. Such work requires adherence to population equality among districts, especially for Congress, where absolute equality of population is almost always required. It must also be precise enough so that the districts comply with the Voting Rights Act, most particularly when localities may be required to draw so-called majority-minority districts, and also accurate enough so that one can determine whether voting in certain jurisdictions is racially polarized.

Having worked in a number of redistricting cases at the level of Congress down to the Village level and having used the Census data in many research projects and in over 100 court cases of various kinds, I decided to do an informal assessment of the new Census data release’s “fitness for redistricting.” I examined the data from Alabama (which brought the case against the inaccurate data) from the latest demonstration product and compared it with the 2010 Census. The results are not encouraging and raise serious questions about the accuracy of the data for redistricting, as well as how the data will be viewed when released to communities across the United States for redistricting and other purposes.

First, at the block level, the important level for redistricting that is used to draw lines and assess plans, the data seems to have been massively changed. In effect, massive numbers of people have been moved from block to block. The table below compares the 2010 data



with the demonstration data, and as one can quickly see, there are substantial changes at the block level:

Population Changes at the Block Level: Most Recent 2010 Census Demonstration Product					
	Total Count	Hispanic	Non-Hispanic White	Non-Hispanic Black	Non-Hispanic Am. Indian Alaskan Native
<b>Totals for Demo Product</b>	4,779,736	185,629	3,204,357	1,244,433	25,903
<b>Totals for 2010 Census</b>	4,779,736	185,602	3,204,402	1,244,437	25,907
<b>State Diff Demo from Census</b>	0	27	-45	-4	-4
<b>Total Increases in Blocks</b>	165,264	53,713	104,550	68,677	13,520
<b>Total Decreases in Blocks</b>	165,264	53,740	104,505	68,673	13,516
<b>Total Changes in Blocks</b>	330,528	107,453	209,055	137,350	27,036
<b>Percent Changes in Blocks</b>	6.92%	57.89%	6.52%	11.04%	104.36%

	Non-Hispanic Asian	Non-Hispanic Nat Ha and Other PI	Non-Hispanic Other	Non-Hispanic 2 or more races
<b>Totals for Demo Product</b>	52,933	1,981	4,027	60,473
<b>Totals for 2010 Census</b>	52,937	1,976	4,030	60,445
<b>State Diff Demo from Census</b>	-4	5	-3	28
<b>Total Increases in Blocks</b>	15,062	1,649	3,317	37,974
<b>Total Decreases in Blocks</b>	15,058	1,654	3,314	38,002
<b>Total Changes in Blocks</b>	30,120	3,303	6,631	75,976
<b>Percent Changes in Blocks</b>	56.90%	167.16%	164.54%	125.69%

All analyses shown in this blog post were performed using the data from the Census Bureau Demo Product released on April 28, 2021, with total epsilon of 12.3, as aggregated and combined with the 2010 Census released data provided by the NHGIS project at the [www.nhgis.org](http://www.nhgis.org). The maps were based upon the Census Boundary files as provided by the Census Bureau.

The total state population was the same for both the demo product and the 2010 Census and the other characteristics were virtually identical at the statewide level. But when one examines the blocks, we can see how much change was induced. For instance, looking at the approximately 1.2 million non-Hispanic Black population in Alabama, about 69,000 were added to some blocks and a similar number were subtracted from other blocks, making for a total of about 11 percent of the Black population having in effect been moved around in Alabama.

Obviously, such moves within the whole state cancel each other out, but that does not answer the question how this would have affected redistricting if these had been the available in 2010. This comparison reveals the important block level changes that



occurred when the Census Bureau used its new technique. For 2020, of course, if the data



occurred when the census bureau used its new technique. For 2020, of course, if the data

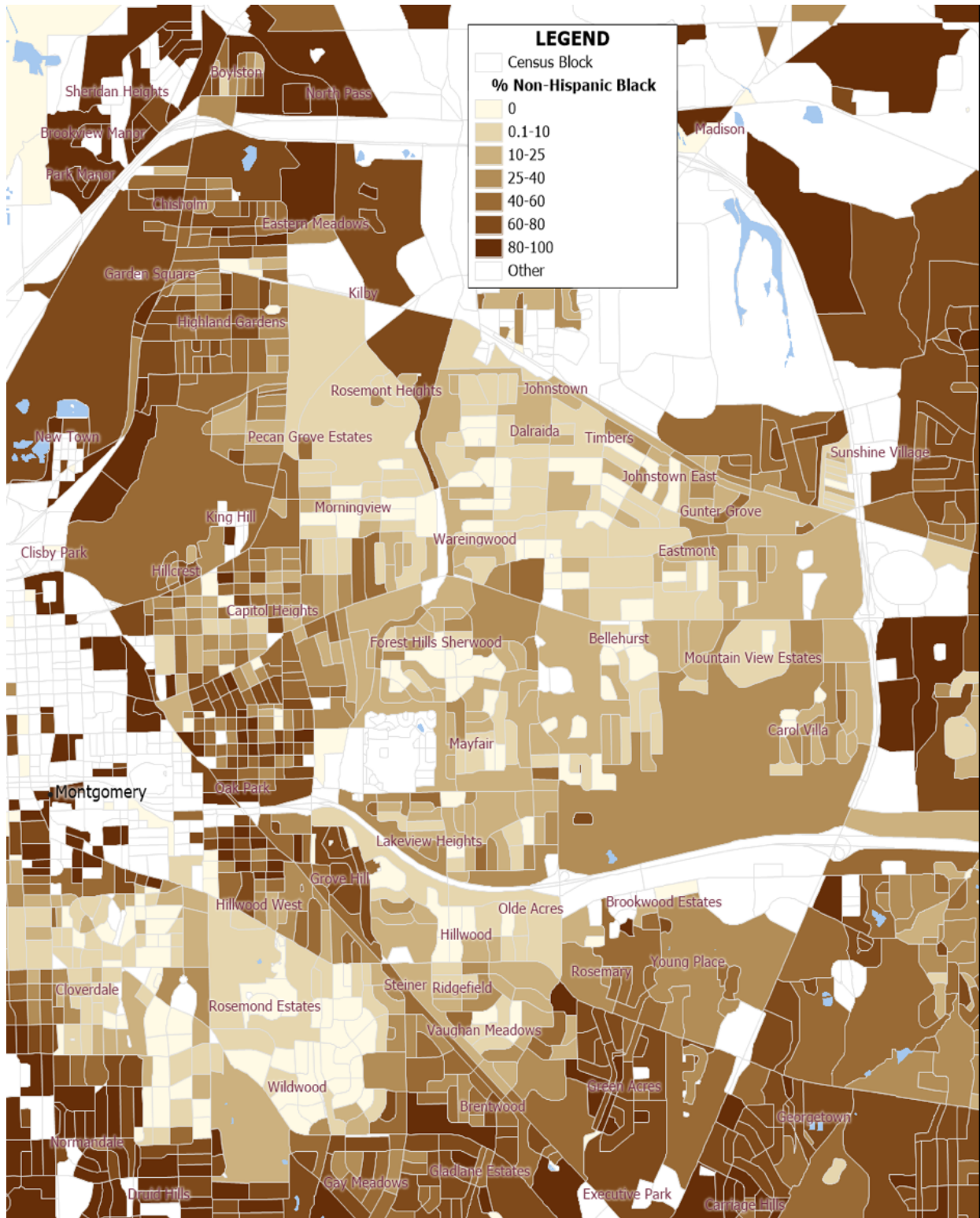
are released with these block-level inaccuracies, we would not know how any plan or district, not to mention blocks, places or other geographies, would be affected, because the manipulated data would be all that was available.

For this reason, I decided to examine the situation of African-Americans in Alabama based upon these data. Map 1 shows the distribution of the non-Hispanic Black population from the 2010 Census in the Montgomery, Ala., area. This is an area with a relatively high Black population concentration. The next two maps show the same area using the demonstration product data, displaying those blocks that have an increase in their non-Hispanic Black population and those blocks that have a decrease (See Maps 2 and 3). In this area (as in all of Alabama) and presumably all the United States, much rearranging of the African-American population occurred. How exactly this would affect redistricting, especially in terms of efforts to protect the voting rights of African-Americans, could not be determined in the time available since the data have only been released for a few days. However, the potential for serious effects is obvious. If the Black population were shifted one way, it might increase their likelihood of being able to argue for an African-American district; if the Black population were shifted another way, it might undermine their claim. But if the 2020 data are released using the new technique, there would be no way to determine which groups were actually eligible for majority-minority districts.

It seems likely that the various minority groups would be spread out more and might make it harder for them to garner a so-called majority-minority seat. This could easily affect redistricting at all levels.

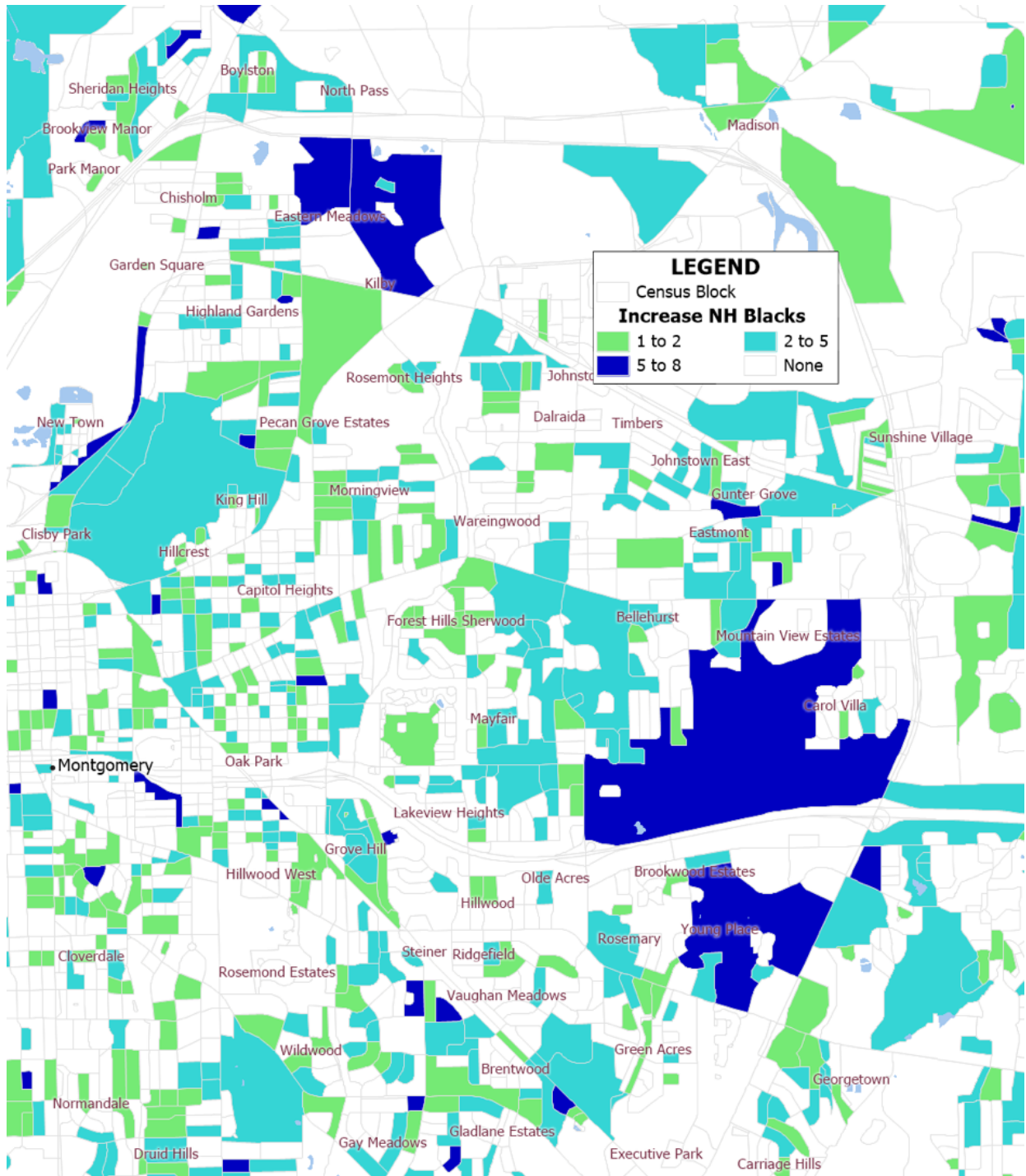
### **Map 1. Non-Hispanic Black Population at the Block Level from 2010 Census**





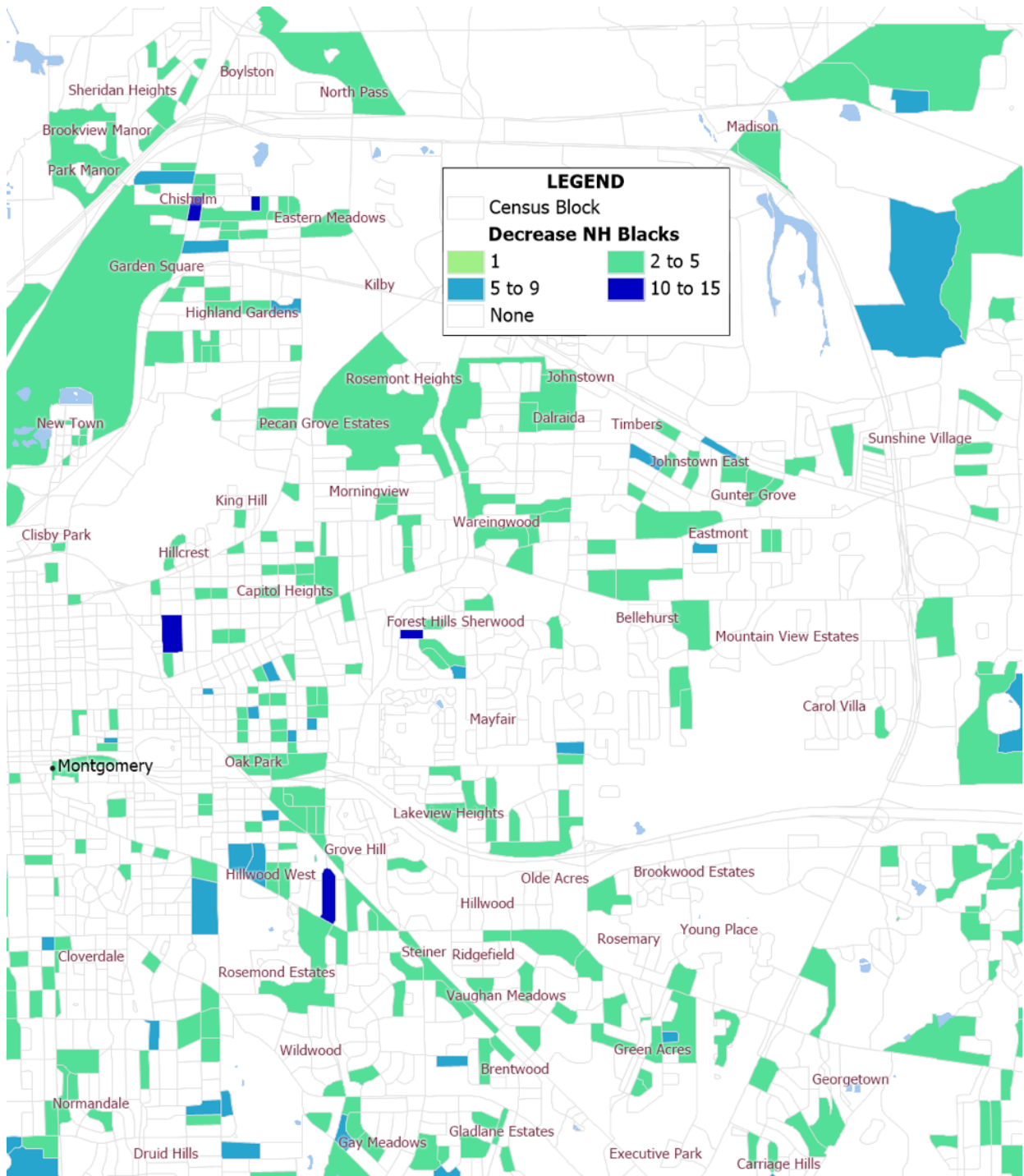
**Map 2. Increases in Non-Hispanic Black Population by Block from Demo Product Compared to 2010 Census as released.**

Compared to 2010 Census as released



**Map 3. Decreases in Non-Hispanic Black Population by Block from Demo Product Compared to 2010 Census as released**





Sparking added concerns is the fact that more 19,000 of the 74,000 blocks with no non-Hispanic Black population in 2010 had Black population added in the demonstration product (See Map 4). For Hispanics, some 29,200 of the 107,000 blocks with no Hispanic



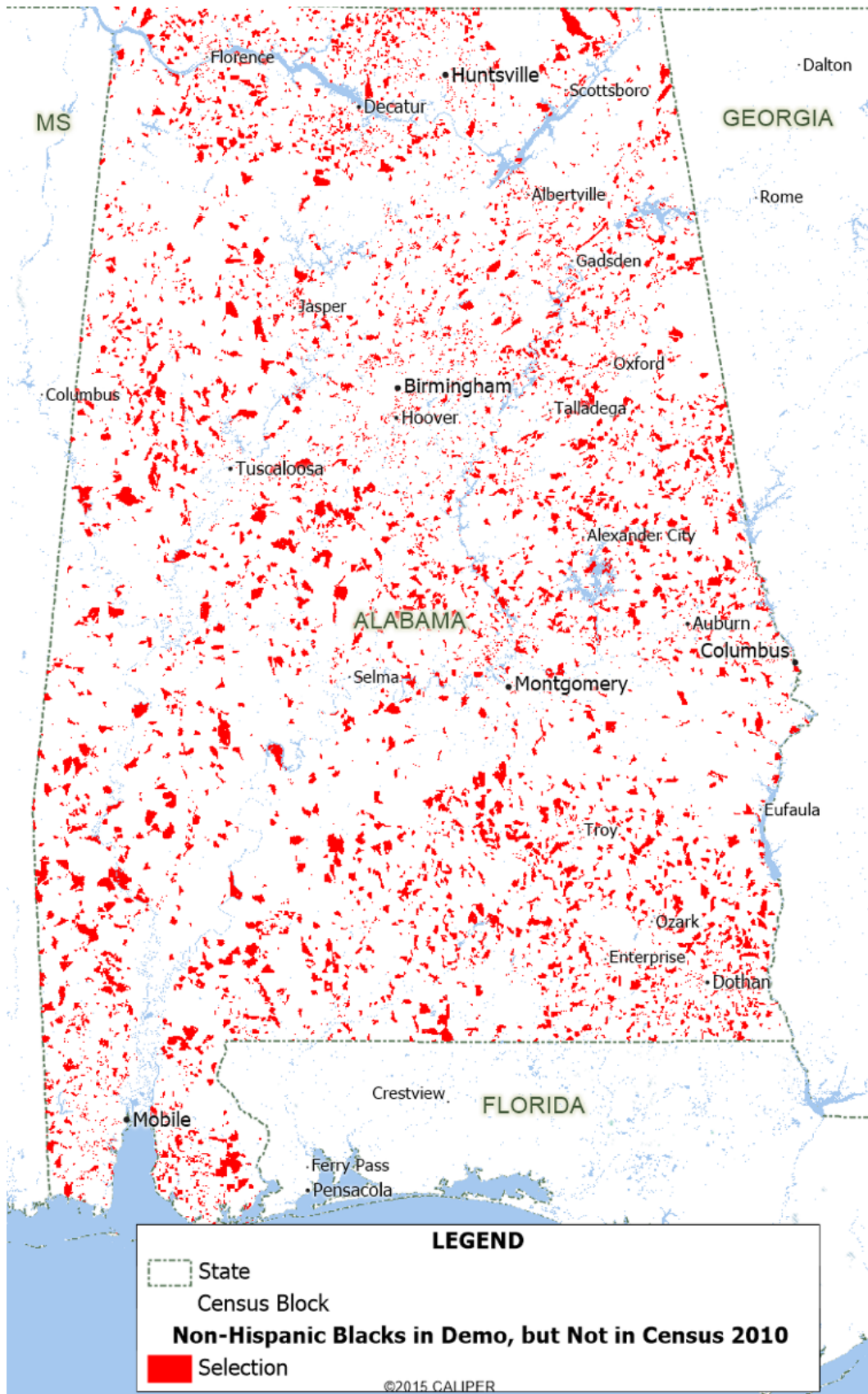
but with other population had Hispanics added to them by the Census process. In short

but with other population had hispanics added to them by the census process. In short,

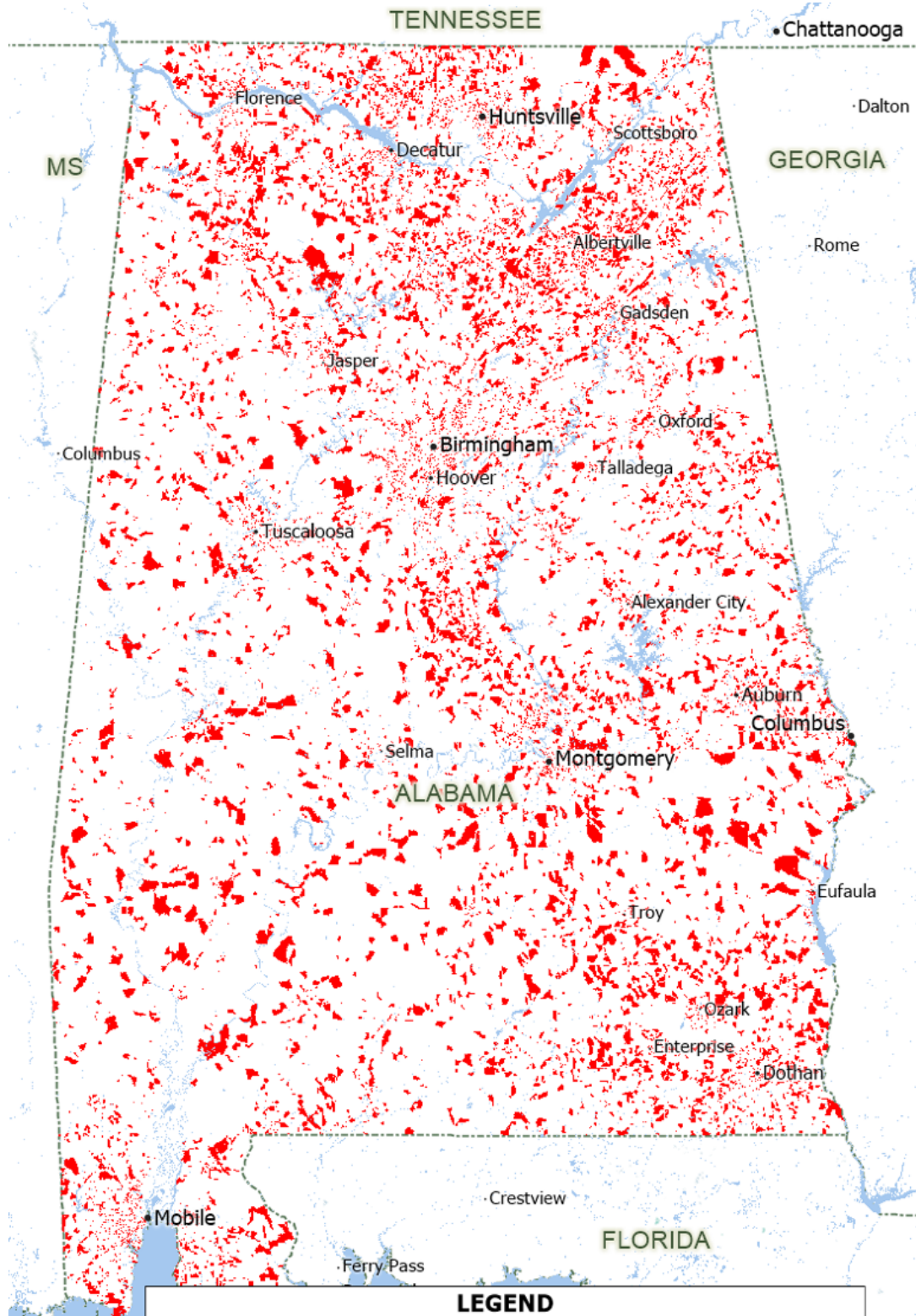
the new Census method of manipulating data in the name of privacy protection generally seems to run a high risk of complicating the difficult task of redistricting, and, perhaps, makes it more difficult for minorities to achieve the goal of a seat that can “elect a candidate of choice,” a major provision of the Voting Rights Act.

#### **Map 4. Blocks in Alabama with No Non-Hispanic Blacks in 2010 Census with One or More Non-Hispanic Blacks in Demo Product**





### Map 5. Blocks in Alabama with No Hispanics in 2010 Census that Had One or More Hispanics in Demo Product





## Conclusions

The latest demonstration product plainly shows the difficulties using it for redistricting. The problems from this admittedly very quick analysis are many, and include the following:

1. The likely difficulty in being able to know where the percent of a minority group is very close to achieving a majority in a district. It appears that the methods have the effect of moving population to less-concentrated areas, which would mean that some districts that seem to not have achieved minority-majority may — with the accurate data — have done so.
2. Complications in analyzing racially polarized voting. The same issues of making the data inaccurate at the block level mean that the denominator used for the two methods of demonstrating racially polarized voting (homogenous precinct analysis and the various regression or regression-like methods) are seriously compromised. In short, if a precinct denominator is not accurate, then the racially polarized voting analysis will be compromised. The degree of the compromise will not be possible to glean from the released data.
3. Because of the two issues enumerated above and others, it will be harder to choose exactly how to draw lines and exactly which areas should be included or excluded in each district. It is also possible that in some cases (one will not know which ones), the total population of a district or a plan will meet or fail to meet various population equality thresholds. It will also be difficult to trade off the various criteria for acceptable plans since all the typical numeric thresholds will not be computed accurately.
4. It will make sharing plans difficult for those overseeing their drawing, as well as various community stakeholders. Displaying maps using blocks that do not





accurately portray the ground truth can only lead to confusion and, perhaps, controversy and distrust.

5. Finally, the new Census Bureau method seems to have other problems, including an inability to create tables that include both persons and households within them (*e.g.*, number and age of persons 17 or younger in a household, tables that are scheduled for the release after PL-94-171).

Given these inaccuracies and limitations, whether this planned method is “ready for prime time” seems very questionable. What is not in question is the large number of states and stakeholders that have tried to alert Census Bureau leadership about the extreme risk that applying this untried method holds for redistricting and other purposes that depends upon the Census for accurate and reliable data.

*(The opinions expressed in this article are those of Andrew Beveridge and may not represent the views of Social Explorer. [His bio is available here.](#))*

---

*Author: Andy Beveridge*

[Back to all posts](#)

**Data insights are waiting to be uncovered**

[Get started](#)

Already using Social Explorer? [Log in.](#)



---

**Product**



---

**Company**



---

**Legal**



---

**Edu Institutions**



---

**Contact Us**

info@socialexplorer.com

(888) 636 - 1118

©2021 Social Explorer



**ARIZONA INDEPENDENT REDISTRICTING COMMISSION**  
*Assorted Neutral Materials Discussing Differential Privacy*

## Table of Contents

Alexandra Wood, et al., <i>Differential Privacy: A Primer for a Non-Technical Audience</i> , 21 VAND. J. ENT. & TECH. L. 209 (Fall 2018).....	464
Samantha Petti, et al., <i>Differential Privacy in the 2020 US Census: What Will It Do? Quantifying the Accuracy/Privacy Tradeoff</i> (Apr. 6, 2020) .....	511
Cynthia Dwork, <i>A Firm Foundation for Private Data Analysis</i> .....	537
Jeffrey Mervis, <i>Can a Set of Equations Keep U.S Census Data Private?</i> , SCIENCE (Jan. 4, 2019) .....	545
Laura McKenna, <i>Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Census of Population and Housing</i> , U.S. CENSUS BUREAU (Nov. 2018) .....	557
Erica Klarreich, <i>Privacy by the Numbers: A New Approach to Safeguarding Data</i> , SCIENTIFIC AMERICAN (Dec. 31, 2012).....	596
Larry Wasserman et al., <i>A Statistical Framework for Differential Privacy</i> (Oc. 22, 2018) .....	611

## 21 Vand. J. Ent. &amp; Tech. L. 209

Vanderbilt Journal of Entertainment and Technology Law  
Fall, 2018

## Article

Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, Salil Vadhan<sup>a1</sup>

Copyright © 2018 by Vanderbilt Journal of Entertainment & Technology Law, Vanderbilt Law School; Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, Salil Vadhan

## DIFFERENTIAL PRIVACY: A PRIMER FOR A NON-TECHNICAL AUDIENCE

## Abstract

*Differential privacy is a formal mathematical framework for quantifying and managing privacy risks. It provides provable privacy protection against a wide range of potential attacks, including those \*210 currently unforeseen. Differential privacy is primarily studied in the context of the collection, analysis, and release of aggregate statistics. These range from simple statistical estimations, such as averages, to machine learning. Tools for differentially private analysis are now in early stages of implementation and use across a variety of academic, industry, and government settings. Interest in the concept is growing among potential users of the tools, as well as within legal and policy communities, as it holds promise as a potential approach to satisfying legal requirements for privacy protection when handling personal information. In particular, differential privacy may be seen as a technical solution for analyzing and sharing data while protecting the privacy of individuals in accordance with existing legal or policy requirements for de-identification or disclosure limitation.*

*This primer seeks to introduce the concept of differential privacy and its privacy implications to non-technical audiences. It provides a simplified and informal, but mathematically accurate, description of differential privacy. Using intuitive illustrations and limited mathematical formalism, it discusses the definition of differential privacy, how differential privacy addresses privacy risks, how differentially private analyses are constructed, and how such analyses can be used in practice. A series of illustrations is used to show how practitioners and policymakers can conceptualize the guarantees provided by differential privacy. These illustrations are also used to explain related concepts, such as composition (the accumulation of risk across multiple analyses), privacy loss parameters, and privacy budgets. This primer aims to provide a foundation that can guide future decisions when analyzing and sharing statistical data about individuals, informing individuals about the privacy protection they will be afforded, and designing policies and regulations for robust privacy protection.*

## Table of Contents

Executive Summary	211
I. Introduction	214
A. Introduction to Legal and Ethical Frameworks for Data Privacy	215
B. Traditional Statistical Disclosure Limitation Techniques	217
C. The Emergence of Formal Privacy Models	218
II. Privacy: A Property of the Analysis--Not Its Output	221
III. What Is the Differential Privacy Guarantee?	225
A. Examples Illustrating What Differential Privacy Protects	227
B. Examples Illustrating What Differential Privacy Does Not Protect	230
IV. How Does Differential Privacy Limit Privacy Loss?	232
A. Differential Privacy and Randomness	233

<i>B. The Privacy Loss Parameter</i>	234
<i>C. Bounding Risk</i>	237
1. A Baseline: Gertrude's Opt-Out Scenario	238
2. Reasoning About Gertrude's Risk	239
<i>D. A General Framework for Reasoning About Privacy Risk</i>	240
<i>E. Composition</i>	244
V. What Types of Analyses Are Performed with Differential Privacy?	246
VI. Practical Considerations When Using Differential Privacy	250
<i>A. The "Privacy Budget"</i>	251
<i>B. Accuracy</i>	254
<i>C. Complying with Legal Requirements for Privacy Protection</i>	259
VII. Tools for Differentially Private Analysis	266
<i>A. Government and Commercial Applications of Differential Privacy</i>	267
<i>B. Research and Development Towards Differentially Private Tools</i>	268
<i>C. Tools for Specific Data Releases or Specific Algorithms</i>	269
VIII. Summary	270
Appendix A. Advanced Topics	271
<i>A.1 How Are Differentially Private Analyses Constructed?</i>	272
<i>A.2 Two Sources of Error: Sampling Error and Added Noise</i>	273
<i>A.3 Group Privacy</i>	275

### \*211 Executive Summary

Differential privacy is a strong, mathematical definition of privacy in the context of statistical and machine learning analysis. It is used to enable the collection, analysis, and sharing of a broad range of statistical estimates based on personal data, such as averages, contingency tables, and synthetic data, while protecting the privacy of the individuals in the data.

\*212 Differential privacy is not a single tool, but rather a criterion, which many tools for analyzing sensitive personal information have been devised to satisfy. It provides a mathematically provable guarantee of privacy protection against a wide range of *privacy attacks*, defined as attempts to learn private information specific to individuals from a data release. Privacy attacks include re-identification, record linkage, and differencing attacks, but may also include other attacks currently unknown or unforeseen. These concerns are separate from security attacks, which are characterized by attempts to exploit vulnerabilities in order to gain unauthorized access to a system.

Computer scientists have developed a robust theory for differential privacy over the last fifteen years, and major commercial and government implementations are starting to emerge.

**The differential privacy guarantee** (Part III). Differential privacy mathematically guarantees that anyone viewing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis.

**The privacy loss parameter** (Section IV.B). What can be learned about an individual as a result of her private information being included in a differentially private analysis is limited and quantified by a privacy loss parameter, usually denoted epsilon ( $\epsilon$ ). Privacy loss can grow as an individual's information is used in multiple analyses, but the increase is bounded as a known function of  $\epsilon$  and the number of analyses performed.

**Interpreting the guarantee** (Section VI.C). The differential privacy guarantee can be understood in reference to other privacy concepts:

- Differential privacy protects an individual's information essentially as if her information were not used in the analysis at all, in the sense that the outcome of a differentially private algorithm is approximately the same whether the individual's information was used or not.

- Differential privacy ensures that using an individual's data will not reveal essentially any personally identifiable information that is specific to her, or even whether the individual's information was used at all. Here, *specific* refers to information that cannot be inferred unless the individual's information is used in the analysis.

As these statements suggest, differential privacy is a new way of protecting privacy that is more quantifiable and comprehensive than the concepts of privacy underlying many existing laws, policies, and practices around privacy and data protection. The differential privacy \*213 guarantee can be interpreted in reference to these other concepts, and can even accommodate variations in how they are defined across different laws. In many settings, data holders may be able to use differential privacy to demonstrate that they have complied with applicable legal and policy requirements for privacy protection.

**Differentially private tools** (Part VII). Differential privacy is currently in initial stages of implementation and use in various academic, industry, and government settings, and the number of practical tools providing this guarantee is continually growing. Multiple implementations of differential privacy have been deployed by corporations such as Google, Apple, and Uber, as well as federal agencies like the US Census Bureau. Additional differentially private tools are currently under development across industry and academia.

Some differentially private tools utilize an interactive mechanism, enabling users to submit queries about a dataset and receive corresponding differentially private results, such as custom-generated linear regressions. Other tools are non-interactive, enabling static data or data summaries, such as synthetic data or contingency tables, to be released and used.

In addition, some tools rely on a curator model, in which a database administrator has access to and uses private data to generate differentially private data summaries. Others rely on a local model, which does not require individuals to share their private data with a trusted third party, but rather requires individuals to answer questions about their own data in a differentially private manner. In a local model, each of these differentially private answers is not useful on its own, but many of them can be aggregated to perform useful statistical analysis.

**Benefits of differential privacy** (Part VIII). Differential privacy is supported by a rich and rapidly advancing theory that enables one to reason with mathematical rigor about privacy risk. Adopting this formal approach to privacy yields a number of practical benefits for users:

- Systems that adhere to strong formal definitions like differential privacy provide protection that is robust to a wide range of potential privacy attacks, including attacks that are unknown at the time of deployment. An analyst using differentially private tools need not anticipate particular types of privacy attacks, as the guarantees of differential privacy hold regardless of the attack method that may be used.
- Differential privacy provides provable privacy guarantees with respect to the cumulative risk from successive data \*214 releases and is the only existing approach to privacy that provides such a guarantee.
- Differentially private tools also have the benefit of transparency, as it is not necessary to maintain secrecy around a differentially private computation or its parameters. This feature distinguishes differentially private tools from traditional de-identification techniques, which often conceal the extent to which the data have been transformed, thereby leaving data users with uncertainty regarding the accuracy of analyses on the data.

- Differentially private tools can be used to provide broad, public access to data or data summaries while preserving privacy. They can even enable wide access to data that cannot otherwise be shared due to privacy concerns. An important example is the use of differentially private synthetic data generation to produce public-use microdata.

Differentially private tools can, therefore, help enable researchers, policymakers, and businesses to analyze and share sensitive data, while providing strong guarantees of privacy to the individuals in the data.

**Keywords:** differential privacy, data privacy, social science research

## I. Introduction

Businesses, government agencies, and research institutions often use and share data containing sensitive or confidential information about individuals.<sup>1</sup> Improper disclosure of such data can have adverse consequences for a data subject's reputation, finances, employability, and insurability, as well as lead to civil liability, criminal penalties, or physical or emotional injuries.<sup>2</sup> Due to these issues and other related concerns, a large body of laws, regulations, ethical codes, institutional policies, contracts, and best practices has emerged to address potential privacy-related harms associated with the collection, use, and release of personal information.<sup>3</sup> The following discussion \*215 provides an overview of the broader data privacy landscape that has motivated the development of formal privacy models like differential privacy.

### A. Introduction to Legal and Ethical Frameworks for Data Privacy

The legal framework for privacy protection in the United States has evolved as a patchwork of highly sector- and context-specific federal and state laws.<sup>4</sup> For instance, Congress has enacted federal information privacy laws to protect certain categories of personal information found in health,<sup>5</sup> education,<sup>6</sup> financial,<sup>7</sup> and government records,<sup>8</sup> among others. These laws often expressly protect information classified as personally identifiable information (PII), which generally refers to information that can be linked to an individual's identity or attributes.<sup>9</sup> Some laws also incorporate de-identification provisions, which provide for the release of information that has been stripped of PII.<sup>10</sup> State data protection and breach notification laws prescribe specific data security and breach reporting requirements when managing certain types of personal information.<sup>11</sup>

In addition, federal regulations generally require researchers conducting studies involving human subjects to secure approval from an institutional review board and fulfill ethical obligations to the participants, such as disclosing the risks of participation, obtaining their informed consent, and implementing specific measures to protect \*216 privacy.<sup>12</sup> It is also common for universities and other research institutions to adopt policies that require their faculty, staff, and students to abide by certain ethical and professional responsibility standards and set forth enforcement procedures and penalties for mishandling data.<sup>13</sup>

Further restrictions apply when privacy-sensitive data are shared under the terms of a data sharing agreement, which will often strictly limit how the recipient can use or redisclose the data received.<sup>14</sup> Organizations may also require privacy measures set forth by technical standards, such as those specifying information security controls to protect personally identifiable information.<sup>15</sup>

In addition, laws such as the EU General Data Protection Regulation are in place to protect personal data about European citizens regardless of where the data reside.<sup>16</sup> International privacy guidelines, such as the privacy principles developed by



the Organisation for Economic Co-operation and Development, have also been adopted by governments across the world.<sup>17</sup> Moreover, the right to privacy is also protected by various international treaties and national constitutions.<sup>18</sup>

Taken together, the safeguards required by these legal and ethical frameworks are designed to protect the privacy of individuals and ensure they fully understand both the scope of personal information to be collected and the associated privacy risks. They also help data holders avoid administrative, civil, and criminal penalties, as well as maintain the public's trust and confidence in commercial, government, and research activities involving personal data.

### *\*217 B. Traditional Statistical Disclosure Limitation Techniques*

A number of technical measures for disclosing data while protecting the privacy of individuals have been produced within the context of these legal and ethical frameworks.<sup>19</sup> In particular, statistical agencies, data analysts, and researchers have widely adopted a collection of statistical disclosure limitation (SDL) techniques to analyze and share data containing privacy-sensitive data with the aim of making it more difficult to learn personal information pertaining to an individual.<sup>20</sup> This category of techniques encompasses a wide range of methods for suppressing, aggregating, perturbing, and generalizing attributes of individuals in the data.<sup>21</sup> Such techniques are often applied with the explicit goal of de-identification-- namely, making it difficult to link an identified person to a record in a data release by redacting or coarsening data.<sup>22</sup>

Advances in analytical capabilities, increases in computational power, and the expanding availability of personal data from a wide range of sources are eroding the effectiveness of traditional SDL techniques.<sup>23</sup> Since the 1990s--and with increasing frequency--privacy and security researchers have demonstrated that data that have been de-identified can often be successfully re-identified via a technique such as record linkage.<sup>24</sup> Re-identification via record linkage, or a linkage attack, refers to the re-identification of one or more records in a de-identified dataset by uniquely linking a record in a de-identified dataset with identified records in a publicly available dataset, such as a voter registration list.<sup>25</sup> As described in Example 1 below, in the late 1990s, Latanya Sweeney famously applied such an attack on a dataset containing de-identified hospital records.<sup>26</sup> Sweeney observed that records in the de-identified dataset contained the date of birth, sex, and \*218 ZIP code of patients; that many of the patients had a unique combination of these three attributes; and that these three attributes were listed alongside individuals' names and addresses in publicly available voting records.<sup>27</sup> Sweeney used this information to re-identify records in the de-identified dataset.<sup>28</sup> Subsequent attacks on protected data have demonstrated weaknesses in other traditional approaches to privacy protection, and understanding the limits of these traditional techniques is the subject of ongoing research.<sup>29</sup>

### *C. The Emergence of Formal Privacy Models*

Re-identification attacks are becoming increasingly sophisticated over time, as are other types of attacks that seek to infer characteristics of individuals based on information about them in a data set.<sup>30</sup> Successful attacks on de-identified data illustrate that traditional technical measures for privacy protection may be particularly vulnerable to attacks devised after a technique's deployment and use.<sup>31</sup> Some de-identification techniques, for example, require the specification of attributes in the data as identifying (e.g., names, dates of birth, or addresses) or non-identifying (e.g., movie ratings or hospital admission dates).<sup>32</sup> Data providers may later discover that attributes initially believed to be non-identifying can in fact be used to re-identify individuals.<sup>33</sup> Similarly, de-identification procedures may require a careful analysis of present and future data sources that could potentially be linked with the de-identified data and enable re-identification of the data. Anticipating the types of attacks and resources an attacker could leverage is a challenging exercise and ultimately will fail to address all potential attacks, as unanticipated \*219 sources of auxiliary information that can be used for re-identification may become available in the future.<sup>34</sup>

Issues such as these underscore the need for privacy technologies that are immune not only to linkage attacks, but to any potential attack, including those currently unknown or unforeseen.<sup>35</sup> They also demonstrate that privacy technologies must provide meaningful privacy protection in settings where extensive external information may be available to potential attackers, such as employers, insurance companies, relatives, and friends of an individual in the data.<sup>36</sup> Real-world attacks further illustrate that ex post remedies, such as simply “taking the data back” when a vulnerability is discovered, are ineffective because many copies of a set of data typically exist, and copies often persist online indefinitely.<sup>37</sup>

In response to the accumulated evidence of weaknesses with respect to traditional approaches, a new privacy paradigm has emerged from the computer science literature--differential privacy.<sup>38</sup> Differential privacy is primarily studied in the context of the collection, analysis, and release of aggregate statistics. Such analyses range from simple statistical estimations-- such as averages--to machine learning.<sup>39</sup> Contrary to common intuition, aggregate statistics such as these are not always safe to release because, as Part III explains, they can often be combined to reveal sensitive information about individual data subjects.

First presented in 2006,<sup>40</sup> differential privacy is the subject of ongoing research to develop privacy technologies that provide robust protection against a wide range of potential attacks.<sup>41</sup> Importantly, differential privacy is not a single tool but a definition or standard for \*220 quantifying and managing privacy risks for which many technological tools have been devised.<sup>42</sup> Analyses performed with differential privacy differ from standard statistical analyses--such as the calculation of averages, medians, and linear regression equations--in that random noise<sup>43</sup> is added in the computation.<sup>44</sup> Tools for differentially private analysis are now in early stages of implementation and use across a variety of academic, industry, and government settings.<sup>45</sup>

This Article provides a simplified and informal, yet mathematically accurate, description of differential privacy.<sup>46</sup> Using intuitive illustrations and limited mathematical formalism, it describes the definition of differential privacy, how it addresses privacy risks, how differentially private analyses are constructed, and how such analyses can be used in practice. This discussion intends to help non-technical audiences understand the guarantees provided by differential privacy. It can help guide practitioners as they make decisions regarding whether to use differential privacy and, if so, what types of promises they should make to data subjects about the guarantees differential privacy provides. In addition, these illustrations intend to help legal scholars and policymakers consider how current and future legal frameworks and instruments should apply to tools based on formal privacy models such as differential privacy.

## \*221 II. Privacy: A Property of the Analysis--Not Its Output

This Article seeks to explain how data containing personal information can be shared in a form that ensures the privacy of the individuals in the data will be protected. The formal study of privacy in the theoretical computer science literature has yielded insights into this problem and revealed why so many traditional privacy-preserving techniques have failed to adequately protect privacy in practice. First, many traditional approaches to privacy failed to acknowledge that attackers could use information obtained from outside the system (i.e., auxiliary information) in their attempts to learn private individual information from a data release.<sup>47</sup> As the amount of detailed auxiliary information continues to grow and become more widely available over time, any privacy-preserving method must take auxiliary information into account in order to provide a reasonable level of privacy protection in light of any auxiliary information that an attacker may hold.<sup>48</sup> Furthermore, traditional approaches treated privacy as a property of the output of an analysis, whereas it is now understood that privacy should be viewed as a property of the analysis itself.<sup>49</sup> Any privacy-preserving method-- including differential privacy--must adhere to this general principle in order to guarantee privacy protection.

The following discussion provides an intuitive explanation of these principles, beginning with a cautionary tale about the re-identification of anonymized records released by the Massachusetts Group Insurance Commission.<sup>50</sup>

**Example 1**

In the late 1990s, the Group Insurance Commission, an agency providing health insurance to Massachusetts state employees, allowed researchers to access anonymized records summarizing information about all hospital visits made by state employees. The agency anticipated that the analysis of these records would lead to recommendations for improving healthcare and controlling <sup>51</sup> healthcare costs.

Massachusetts Governor William Weld reassured the public that steps would be taken to protect the privacy of patients in the data. Before releasing the records to researchers, the agency removed names, addresses, Social Security numbers, and other pieces of information that could be used to identify individuals in the records.

Viewing this as a challenge, Professor Latanya Sweeney, then a graduate student at MIT, set out to identify Governor Weld's record in the dataset. She obtained demographic information about Governor Weld, including his ZIP code and date of birth, by requesting a copy of voter registration records made available to the public for a small fee. Finding just one record in the anonymized medical claims dataset that matched Governor Weld's gender, ZIP code, and date of birth enabled her to mail the Governor a copy of his personal medical records.

As Example 1 illustrates, in many cases, a dataset that appears to be anonymous may nevertheless be used to learn sensitive information about individuals. In her demonstration, Professor Sweeney used voter registration records as auxiliary information in an attack. This re-identification demonstrates the importance of using privacy-preserving methods that are robust to auxiliary information that may be exploited by an adversary. Following Professor Sweeney's famous demonstration, a long series of attacks has been carried out against different types of data releases anonymized using a wide range of techniques and auxiliary information.<sup>51</sup> These attacks have shown that risks remain even if additional pieces of information, such as those that were leveraged in Professor Sweeney's attack (gender, date of birth, and ZIP code), are removed from a dataset prior to release.<sup>52</sup> Risks also remain when using some traditional SDL techniques, such as  $k$ -anonymity, which is satisfied for a dataset in which the identifying attributes that appear for each person are identical to those of at least  $k - 1$  other individuals in the dataset.<sup>53</sup> Research has continually demonstrated that privacy measures that treat privacy as a property of <sup>54</sup> the output, such as  $k$ -anonymity and other traditional statistical disclosure limitation techniques, will fail to protect privacy.

The Authors offer a brief note on terminology before proceeding. The discussions throughout this Article use the terms “analysis” and “computation” interchangeably to refer to any transformation, usually performed by a computer program, of input data into some output.

As an example, consider an analysis on data containing personal information about individuals. The analysis may be as simple as determining the average age of the individuals in the data, or it may be more complex and utilize sophisticated modeling and inference techniques. In any case, the analysis involves performing a computation on input data and outputting the result. Figure 1 illustrates this notion of an analysis.

**Figure 1. An Analysis**

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

This primer focuses, in particular, on analyses for transforming sensitive personal data into an output that can be released publicly. For example, an analysis may involve the application of techniques for aggregating or de-identifying a set of personal data in order to produce a sanitized version of the data that is safe to release. The data provider will want to ensure that publishing the output of this computation will not unintentionally leak information from the privacy-sensitive input data--but how?

A key insight from the theoretical computer science literature is that privacy is a property of the informational relationship between the input and output, not a property of the output alone.<sup>54</sup> The following discussion illustrates why this is the case through a series of examples.

### *Example 2*

Anne, a staff member at a high school, would like to include statistics about student performance in a presentation. She \*224 considers publishing the fact that the GPA of a representative ninth-grade student is 3.5. Because the law protects certain student information held by educational institutions, she must ensure that the statistic will not inappropriately reveal student information, such as the GPA of any particular student.

One might naturally think that Anne could examine the statistic itself and determine that it is unlikely to reveal private information about an individual student. However, although the publication of this statistic might seem harmless, Anne needs to know how the statistic was computed to make that determination. For instance, if the representative ninth-grade GPA was calculated by taking the GPA of the alphabetically first student in the school, then the statistic completely reveals the GPA of that student.<sup>55</sup>

### *Example 3*

Alternatively, Anne considers calculating a representative statistic based on average features of the ninth graders at the school. She takes the most common first name, the most common last name, the average age, and the average GPA for the ninth-grade class. What she produces is “John Smith, a fourteen-year-old in the ninth grade, has a 3.1 GPA.” Anne includes this statistic and the method used to compute it in her presentation. In an unlikely turn of events, a new ninth-grade student named John Smith joins the class the following week.

Although the output of Anne's analysis *looks* like it reveals private information about the new ninth grader John Smith, it actually does not-- because the analysis itself was not based on his student records in any way. While Anne might decide to present the statistic differently to avoid confusion, using it would not reveal private information about John. It may seem counterintuitive that releasing a “representative” GPA violates privacy (as shown by Example 2), while releasing a GPA attached to a student's name would not (as shown by Example 3). Yet these examples illustrate that the key to preserving \*225 privacy is the informational relationship between the private input and the public output--and not the output itself. Furthermore, not only is it necessary to examine the analysis itself to determine whether a statistic can be published while preserving privacy, but it is also sufficient. In other words, if one knows whether the process used to generate a statistic preserves privacy, the output statistic does not need to be considered at all.

## III. What Is the Differential Privacy Guarantee?

The previous Part illustrates why privacy should be thought of as a property of a computation--but how does one know whether a particular computation has this property?

Intuitively, a computation protects the privacy of individuals in the data if its output does not reveal any information that is specific to any individual data subject. Differential privacy formalizes this intuition as a mathematical definition.<sup>56</sup> Just as we can show that an integer is even by demonstrating that it is divisible by two, we can show that a computation is differentially private by proving it meets the constraints of the definition of differential privacy. In turn, if a computation can be proven to be differentially private, we can rest assured that using the computation will not unduly reveal information *specific* to any data subject.<sup>57</sup> Here, the term *specific* refers to information that cannot be inferred unless the individual's information is used in

the analysis. For example, the information released by Anne in Example 3 is not specific to the new ninth grader John Smith because it is computed without using his information.

The following example illustrates how differential privacy formalizes this intuitive privacy requirement as a definition.

#### **Example 4**

Researchers have selected a sample of individuals across the United States to participate in a survey exploring the relationship between socioeconomic status and health outcomes. The participants were asked to complete a questionnaire covering topics concerning their residency, their finances, and their medical history.

**\*226** One of the participants, John, is aware that individuals have been re-identified in previous releases of de-identified data and is concerned that personal information he provides about himself, such as his medical history or annual income, could one day be revealed in de-identified data released from this study. If leaked, this information could lead to a higher life insurance premium or an adverse decision with respect to a future mortgage application.<sup>58</sup>

Differential privacy can be used to address John's concerns. If the researchers promise they will only share survey data after processing the data with a differentially private computation, John is guaranteed that any data the researchers release will disclose essentially nothing that is specific to him, even though he participated in the study.<sup>59</sup> To understand what this means, consider the thought experiment, illustrated in Figure 2 and referred to as John's opt-out scenario. In John's opt-out scenario, an analysis is performed using data about the individuals in the study, except that information about John is omitted. His privacy is protected in the sense that the outcome of the analysis does not depend on his specific information-- because his information was not used in the analysis at all.

#### **Figure 2. John's Opt-Out Scenario**

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

John's opt-out scenario differs from the real-world scenario depicted in Figure 1, where John's information is part of the input of the analysis along with the personal information of the other study participants. In contrast to his opt-out scenario, the real-world scenario involves some potential risk to John's privacy. Some of his personal information could **\*227** be revealed by the outcome of the analysis because his information was used as input to the computation.<sup>60</sup>

#### **A. Examples Illustrating What Differential Privacy Protects**

Differential privacy aims to protect John's privacy in the real-world scenario in a way that mimics the privacy protection he is afforded in his opt-out scenario.<sup>61</sup> In other words, what can be learned about John from a differentially private computation is essentially limited to what could be learned about him from everyone else's data without his own data being included in the computation. Crucially, this same guarantee is made not only with respect to John, but also with respect to every other individual contributing her information to the analysis.

A precise description of the differential privacy guarantee requires using formal mathematical language, as well as technical concepts and reasoning that are beyond the scope of this Article. In lieu of the mathematical definition, this Article offers a few illustrative examples to discuss various aspects of differential privacy in a way designed to be intuitive and generally accessible. The scenarios in this Section illustrate the types of information disclosures that are addressed when using differential privacy.

#### **Example 5**

Alice and Bob are professors at Private University. They both have access to a database that contains personal information about students at the university, including information related to the financial aid each student receives. Because it contains personal information, access to the database is restricted. To gain access, Alice and Bob were required to demonstrate they planned to follow the university's protocols for handling personal data by undergoing confidentiality training and signing data use agreements \*228 proscribing their use and disclosure of personal information obtained from the database.

In March, Alice publishes an article based on the information in this database and writes that “the current freshman class at Private University is made up of 3,005 students, 202 of whom are from families earning over \$350,000 per year.” Alice reasons that, because she published an aggregate statistic taken from over 3,005 people, no individual's personal information will be exposed. The following month, Bob publishes a separate article containing these statistics: “201 students in Private University's freshman class of 3,004 have household incomes exceeding \$350,000 per year.” Neither Alice nor Bob is aware that they have both published similar information.

A clever student Eve reads both of these articles and makes an observation. From the published information, Eve concludes that between March and April one freshman withdrew from Private University and that the student's parents earn over \$350,000 per year. Eve asks around and is able to determine that a student named John dropped out around the end of March. Eve then informs her classmates that John's family probably earns over \$350,000 per year.

John hears about this and is upset that his former classmates learned about his family's financial status. He complains to the university, and Alice and Bob are asked to explain. In their defense, both Alice and Bob argue that they published only information that had been aggregated over a large population and does not identify any individuals.

Example 5 illustrates how, in combination, the results of multiple analyses using information about the same people may enable one to draw conclusions about individuals in the data. Alice and Bob each published information that, in isolation, seems innocuous. However, when combined, the information they published compromised John's privacy. This type of privacy breach is difficult for Alice or Bob to prevent individually, as neither knows what information others have already revealed or will reveal in future. This is referred to as the problem of composition.<sup>62</sup>

\*229 Suppose, instead, that the institutional review board at Private University only allows researchers to access student records by submitting queries to a special data portal. This portal responds to every query with an answer produced by running a differentially private computation on the student records. As explained in Part IV, differentially private computations introduce a carefully tuned amount of random noise to the statistics outputted.<sup>63</sup> This means that the computation gives an approximate answer to every question asked through the data portal.<sup>64</sup> As Example 6 illustrates, the use of differential privacy prevents the privacy leakage that occurred in Example 5.

### **Example 6**

In March, Alice queries the data portal for the number of freshmen who come from families with a household income exceeding \$350,000. The portal returns the noisy count of 204, leading Alice to write in her article that “the current freshman class at Private University includes approximately 200 students from families earning over \$350,000 per year.” In April, Bob asks the same question and gets the noisy count of 199 students. Bob publishes in his article that “approximately 200 families in Private University's freshman class have household incomes exceeding \$350,000 per year.” The publication of these noisy figures prevents Eve from concluding that one student, with a household income greater than \$350,000, withdrew from the university in March. The risk that John's personal information could be uncovered based on these publications is thereby reduced.

Example 6 hints at one of the most important properties of differential privacy--it is robust under composition.<sup>65</sup> If multiple analyses are performed on data describing the same set of individuals, then, as long as each of the analyses satisfies differential privacy, it is guaranteed that all of the information released, when taken together, will still be differentially private.<sup>66</sup> Notice how

this example is \*230 markedly different from Example 5, in which Alice and Bob do not use differentially private analyses and inadvertently release two statistics that, when combined, lead to the full disclosure of John's personal information. The use of differential privacy rules out the possibility of such a complete breach of privacy. This is because differential privacy enables one to measure and bound the cumulative privacy risk from multiple analyses of information about the same individuals.<sup>67</sup>

It is important to note, however, that every analysis, regardless of whether it is differentially private or not, results in some leakage of information about the individuals whose information is being analyzed. This is a well-established principle within the statistical community, as evidenced by a 2005 report that concluded "[t]he release of statistical data inevitably reveals some information about individual data subjects."<sup>68</sup> Furthermore, this leakage accumulates with each analysis, potentially to a point where an attacker may infer the underlying data.<sup>69</sup> This is true for every release of data, including releases of aggregate statistics.<sup>70</sup> In particular, releasing too many aggregate statistics too accurately inherently leads to severe privacy loss.<sup>71</sup> For this reason, there is a limit to how many analyses can be performed on a specific dataset while providing an acceptable guarantee of privacy.<sup>72</sup> This is why it is critical to measure privacy loss and to understand quantitatively how risk accumulates across successive analyses, as Sections IV.E and VI.A describe below.

### ***B. Examples Illustrating What Differential Privacy Does Not Protect***

The following examples illustrate the types of information disclosures differential privacy does not seek to address.

#### ***Example 7***

Suppose Ellen is a friend of John's and knows some of his habits, such as that he regularly consumes several glasses of red wine with \*231 dinner. Ellen learns that John took part in a large research study, and that this study found a positive correlation between drinking red wine and the likelihood of developing a certain type of cancer. She might therefore conclude, based on the results of this study and her prior knowledge of John's drinking habits, that he has a heightened risk of developing cancer.

It may seem at first that the publication of the results from the research study enabled a privacy breach by Ellen. After all, learning about the study's findings helped her infer new information about John that he himself may be unaware of (i.e., his elevated cancer risk). However, notice that Ellen would be able to infer this information about John even if John had not participated in the medical study (i.e., it is a risk that exists in both John's opt-out scenario and the real-world scenario).<sup>73</sup> Risks of this nature apply to everyone, regardless of whether they shared personal data through the study or not. Consider another example:

#### ***Example 8***

Ellen knows that her friend John is a public school teacher with five years of experience and that he is about to start a job in a new school district. She later comes across a local news article about a teachers' union dispute, which includes salary figures for the public school teachers in John's new school district. Ellen is able to approximately determine John's salary at his new job, based on the district's average salary for a teacher with five years of experience.

Note that, as in the previous example, Ellen can determine information about John (i.e., his new salary) from the published information, even though the published information was not based on John's information. In both examples, John could be adversely affected by the discovery of the results of an analysis, even in his opt-out scenario. In both John's opt-out scenario and in a differentially private real-world scenario, it is therefore not guaranteed that no information about John can be revealed. The use of differential privacy limits the revelation of information *specific* to John.

\*232 These examples suggest, more generally, that any useful analysis carries a risk of revealing some information about individuals. One might observe, however, that such risks are largely unavoidable. In a world in which data about individuals are

collected, analyzed, and published, John cannot expect better privacy protection than is offered by his opt-out scenario because he has no ability to prevent others from participating in a research study or appearing in public records.

Moreover, the types of information disclosures enabled in John's opt-out scenario often result in individual and societal benefits. For example, the discovery of a causal relationship between red wine consumption and elevated cancer risk can lead to new public health recommendations, support future scientific research, and inform John about possible changes he could make in his habits that would likely have positive effects on his health. Similarly, the publication of public school teacher salaries may be seen as playing a critical role in transparency and public policy, as it can help communities make informed decisions regarding appropriate salaries for their public employees.

#### IV. How Does Differential Privacy Limit Privacy Loss?

The previous Part explains that the only things that can be learned about a data subject from a differentially private data release are essentially what could have been learned if the analysis had been performed without that individual's data.

How do differentially private analyses achieve this goal? And what is meant by “essentially” when stating that the only things that can be learned about a data subject are essentially those things that could be learned without the data subject's information? The answers to these two questions are related. Differentially private analyses protect the privacy of individual data subjects by introducing carefully tuned random noise when producing statistics.<sup>74</sup> Differentially private analyses are also allowed to leak *some* small amount of information specific to individual data subjects.<sup>75</sup> A privacy parameter controls exactly how much information can be leaked and, relatedly, how much random noise is introduced during the differentially private computation.<sup>76</sup>

##### *\*233 A. Differential Privacy and Randomness*

Example 6 shows that differentially private analyses introduce random noise to the statistics they produce. Intuitively, this noise masks the differences between the real-world computation and the opt-out scenario of each individual in the dataset. This means that the outcome of a differentially private analysis is not exact, but rather an approximation. In addition, a differentially private analysis may, if performed twice on the same dataset, return different results because it intentionally introduces random noise. Therefore, analyses performed with differential privacy differ from standard statistical analyses, such as the calculation of averages, medians, and linear regression equations, in which one gets the same answer when a computation is repeated twice on the same dataset.

##### *Example 9*

Consider a differentially private analysis that computes the number of students in a sample with a GPA of at least 3.0. Say that there are 10,000 students in the sample, and exactly 5,603 of them have a GPA of at least 3.0. An analysis that added no random noise would report that 5,603 students had a GPA of at least 3.0.

A differentially private analysis, however, introduces random noise to protect the privacy of the data subjects. For instance, a differentially private analysis might report an answer of 5,521 when run on the student data; when run a second time on the same data, it might report an answer of 5,586.<sup>77</sup>

Although a differentially private analysis might produce many different answers given the same dataset, it is usually possible to calculate accuracy bounds for the analysis measuring how much an output of the analysis is expected to differ from the noiseless answer.<sup>78</sup> Section VI.B discusses how the random noise introduced by a differentially private analysis affects statistical accuracy. Appendix A.1 *\*234* provides more information about the role randomness plays in the construction of differentially private analyses.



### B. The Privacy Loss Parameter

An essential component of a differentially private computation is the privacy loss parameter, which determines how well each individual's information needs to be hidden and, consequently, how much noise needs to be introduced.<sup>79</sup> It can be thought of as a tuning knob for balancing privacy and accuracy. Each differentially private analysis can be tuned to provide more or less privacy--resulting in less or more accuracy, respectively--by changing the value of this parameter. The parameter can be thought of as limiting how much a differentially private computation is allowed to deviate from the opt-out scenario of each individual in the data.

Consider the opt-out scenario for a certain computation, such as estimating the number of HIV-positive individuals in a surveyed population. Ideally, this estimate should remain exactly the same whether or not a single individual, such as John discussed above, is included in the survey. However, as described above, ensuring that the estimate is *exactly* the same would require the total exclusion of John's information from the real-world analysis. It would also require excluding the information of other individuals (e.g., that of Gertrude, Peter, and so forth) in order to provide perfect privacy protection for them as well. Continuing this line of argument, one can conclude that the personal information of every single surveyed individual must be removed in order to satisfy each individual's opt-out scenario. Thus, the analysis cannot rely on any person's information and is completely useless.

To avoid this dilemma, differential privacy requires only that the output of the analysis remain approximately the same, whether John participates in the survey or not. That is, differential privacy allows for a deviation between the output of the real-world analysis and that of each individual's opt-out scenario. A parameter quantifies and limits the extent of the deviation between the opt-out and real-world scenarios.<sup>80</sup> As Figure 3 illustrates below, this parameter is usually denoted by the Greek letter  $\epsilon$  (epsilon) and referred to as the privacy parameter or, more accurately, the privacy loss parameter.<sup>81</sup> The parameter  $\epsilon$  measures the effect of each individual's information on the \*235 output of the analysis. It can also be viewed as a measure of the additional privacy risk an individual could incur beyond the risk incurred in the opt-out scenario. Note that Figure 3 replaces John with an arbitrary individual  $X$  to emphasize that the differential privacy guarantee is made simultaneously to all individuals in the sample--not just John.

#### Figure 3. Differential Privacy

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

Moreover, it can be shown that the deviation between the real-world and opt-out scenarios cannot be increased by any further processing of the output of a differentially private analysis. Hence, the guarantees of differential privacy, described below, hold regardless of how an attacker may try to manipulate the output. In this sense, differential privacy is robust to a wide range of potential privacy attacks, including attacks that are unknown at the time of deployment.<sup>82</sup>

Choosing a value for  $\epsilon$  can be thought of as setting the desired level of privacy protection. This choice also affects the utility or accuracy that can be obtained from the analysis.<sup>83</sup> A smaller value of  $\epsilon$  results in a smaller deviation between the real-world analysis and each opt-out scenario and is therefore associated with stronger privacy \*236 protection but less accuracy.<sup>84</sup> For example, when  $\epsilon$  is set to zero, the real-world differentially private analysis mimics the opt-out scenario of each individual perfectly and simultaneously. However, an analysis that perfectly mimics the opt-out scenario of each individual would require ignoring all information from the input and, accordingly, could not provide any meaningful output. Yet, when  $\epsilon$  is set to a small number such as 0.1, the deviation between the real-world computation and each individual's opt-out scenario will be small, providing strong privacy protection, while also enabling an analyst to derive useful statistics based on the data.

Accepted guidelines for choosing  $\epsilon$  have not yet been developed.<sup>85</sup> The increasing use of differential privacy in real-life applications will likely shed light on how to reach a reasonable compromise between privacy and accuracy, and the accumulated evidence from these real-world decisions will likely contribute to the development of future guidelines.<sup>86</sup> As discussed in Section IV.D, the Authors of this Article recommend that, when possible,  $\epsilon$  be set to a small number, such as a  $\epsilon$  value less than 1.<sup>87</sup> As Figure 3 illustrates, the maximum deviation between the opt-out scenario and the real-world computation should hold simultaneously for each individual  $X$  whose information is included in the input.

### ***C. Bounding Risk***

The previous Section discusses how the privacy loss parameter limits the deviation between the real-world computation and each data subject's opt-out scenario. However, it might not be clear how this abstract guarantee relates to the privacy concerns individuals face in the real world. To help ground the concept, this Section discusses a practical interpretation of the privacy loss parameter. It describes how the parameter can be understood as a bound on the financial risk incurred by an individual participating in a research study.

Any useful analysis carries the risk that it will reveal information about the individuals in the data.<sup>88</sup> An individual whose information is used in an analysis may be concerned that a potential leakage of her personal information could result in reputational, financial, or other costs. Examples 10 and 11 below introduce a scenario in which an individual participating in a research study worries that an analysis on the data collected in the research study may leak information that could lead to a substantial increase in her life insurance premium. Example 12 illustrates that, while differential privacy necessarily cannot fully eliminate this risk, it can guarantee that the risk will be limited by quantitative bounds that depend on  $\epsilon$ .<sup>89</sup>

#### ***Example 10***

Gertrude, a sixty-five-year-old woman, is considering whether to participate in a medical research study. While she can envision many potential personal and societal benefits resulting in part from her participation in the study, she is concerned that the personal information she discloses over the course of the study could lead to an increase in her life insurance premium in the future.

For example, Gertrude is concerned that the tests she would undergo as part of the research study would reveal that she is predisposed to suffer a stroke and is significantly more likely to die  $\epsilon$  in the coming year than the average person of her age and gender. If such information related to Gertrude's increased risk of morbidity and mortality is discovered by her life insurance company, it will likely increase the premium for her annual renewable term policy substantially.

Before she opts to participate in the study, Gertrude wishes to be assured that privacy measures are in place to ensure that her participation will have, at most, a limited effect on her life insurance premium.

#### **1. A Baseline: Gertrude's Opt-Out Scenario**

It is important to note that Gertrude's life insurance company may raise her premium based on something it learns from the medical research study, even if Gertrude does not herself participate in the study. The following example is provided to illustrate such a scenario.<sup>90</sup>

#### ***Example 11***

Gertrude holds a \$100,000 life insurance policy. Her life insurance company has set her annual premium at \$1,000, i.e., 1% of \$100,000, based on actuarial tables showing that someone of Gertrude's age and gender has a 1% chance of dying in the next year.

Suppose Gertrude opts out of participating in the medical research study. Regardless, the study reveals that coffee drinkers are more likely to suffer a stroke than non-coffee drinkers. Gertrude's life insurance company may update its assessment and conclude that, as a sixty-five-year-old woman who drinks coffee, Gertrude has a 2% chance of dying in the next year. The company decides to increase Gertrude's annual premium from \$1,000 to \$2,000 based on the findings of the study.<sup>91</sup>

**\*239** In this example, the results of the study led to an increase in Gertrude's life insurance premium, even though she did not contribute any personal information to the study. A potential increase of this nature is unavoidable to Gertrude in this scenario because she cannot prevent other people from participating in the study. This example illustrates that Gertrude can experience a financial loss even in her opt-out scenario. Because, as presented in this example, Gertrude cannot avoid this type of risk on her own,<sup>92</sup> in the following discussion this opt-out scenario will serve as a baseline for measuring potential increases in her privacy risk above this threshold.

## 2. Reasoning About Gertrude's Risk

Next consider the increase in risk, relative to Gertrude's opt-out scenario, that is due to her participation in the study.

### *Example 12*

Suppose Gertrude decides to participate in the research study. Based on the results of medical tests performed on Gertrude over the course of the study, the researchers conclude that Gertrude has a 50% chance of dying from a stroke in the next year. If the data from the study were to be made available to Gertrude's insurance company, it might decide to increase her insurance premium to \$50,000 in light of this discovery.

Fortunately for Gertrude, this does not happen. Rather than releasing the full dataset from the study, the researchers release only a differentially private summary of the data they collected. Differential privacy guarantees that, if the researchers use a value of  $\epsilon = 0.01$ , then the insurance company's estimate of the probability that Gertrude will die in the next year can increase from the opt-out scenario's estimate of 2% to at most

$$2\% \cdot (1 + 0.01) = 2.02\%.$$

**\*240** Thus Gertrude's insurance premium can increase from \$2,000 to, at most, \$2,020. Gertrude's first-year cost of participating in the research study, in terms of a potential increase in her insurance premium, is at most \$20.

Note that this does not mean that the insurance company's estimate of the probability that Gertrude will die in the next year will necessarily increase as a result of her participation in the study, nor that if the estimate increases it must increase to 2.02%. What the analysis shows is that if the estimate were to increase it would not exceed 2.02%.

In this example, Gertrude is aware of the fact that the study could indicate that her risk of dying in the next year exceeds 1%. She happens to believe, however, that the study will not indicate more than a 2% risk of dying in the next year, in which case the potential cost to her of participating in the research will be at most \$20. Based on her belief, Gertrude may decide that she considers the potential cost of \$20 to be too high and that she cannot afford to participate with this value of  $\epsilon$  and this level of risk. Alternatively, she may decide that it is worthwhile. Perhaps she is paid more than \$20 to participate in the study, or the information she learns from the study is worth more than \$20 to her. The key point is that differential privacy allows Gertrude to make a more informed decision based on the worst-case cost of her participation in the study.

It is worth noting that, should Gertrude decide to participate in the study, her risk might increase--even if her insurance company is not aware of her participation. Gertrude might actually have a higher chance of dying in the next year, and that could affect the study results. In turn, her insurance company might decide to raise her premium because she fits the profile of the studied population--even if it does not believe her data were included in the study. Differential privacy guarantees that, even if the insurance company knows that Gertrude *did* participate in the study--it can only make inferences about her that it could have essentially made if she had not participated in the study.

#### ***D. A General Framework for Reasoning About Privacy Risk***

Gertrude's scenario illustrates how differential privacy is a general framework for reasoning about the increased risk that is incurred when an individual's information is included in a data analysis. Differential privacy guarantees that an individual will be exposed to essentially the same privacy risk, whether or not her data <sup>\*241</sup> are included in a differentially private analysis.<sup>93</sup> In this context, one can think of the privacy risk associated with a release of the output of a data analysis as the potential harm that an individual might incur because of a belief that an observer forms based on that data release.

In particular, when  $\epsilon$  is set to a small value, an observer's posterior belief can change--relative to the case where the data subject is not included in the data set--by a factor of at most approximately  $1 + \epsilon$  based on a differentially private data release.<sup>94</sup> For example, if  $\epsilon$  is set to 0.01, then the privacy risk to an individual resulting from participation in a differentially private computation grows by at most a multiplicative factor of 1.01.

As Examples 11 and 12 illustrate, there is a risk to Gertrude that the insurance company will see the study results, update its beliefs about the mortality of Gertrude, and charge her a higher premium. If the insurance company infers from the study results that Gertrude has probability  $p$  of dying in the next year and her insurance policy is valued at \$100,000, her premium will increase to  $p \times \$100,000$ . This risk exists, even if Gertrude does not participate in the study. Recall how, in Example 11, the insurance company's belief that Gertrude will die in the next year doubles from 1% to 2%, increasing her premium from \$1,000 to \$2,000, based on general information learned from the individuals who did participate. Recall also that if Gertrude does decide to participate in the study (as in Example 12), differential privacy limits the change in this risk relative to her opt-out scenario. In financial terms, her risk increases by at most \$20, since the insurance company's beliefs about her probability of death change from 2% to at most  $2\% \cdot (1 + \epsilon) = 2.02\%$ , where  $\epsilon = 0.01$ .

Note that the above calculation requires certain information that may be difficult to determine in the real world. In particular, the 2% baseline in Gertrude's opt-out scenario (i.e., Gertrude's insurer's belief about her chance of dying in the next year) is dependent on the results from the medical research study, which Gertrude does not know at the time she makes her decision whether to participate. Fortunately, differential privacy provides guarantees relative to every baseline risk.<sup>95</sup>

#### ***\*242 Example 13***

Say that, without her participation, the study results would lead the insurance company to believe that Gertrude has a 3% chance of dying in the next year (instead of the 2% chance hypothesized earlier). This means that Gertrude's insurance premium would increase to \$3,000. Differential privacy guarantees that, if Gertrude had instead decided to participate in the study, the insurer's estimate for Gertrude's mortality would have been at most  $3\% \cdot (1 + \epsilon) = 3.03\%$  (assuming an  $\epsilon$  of 0.01), which means that her premium would not increase beyond \$3,030.

Calculations like those used in the analysis of Gertrude's privacy risk can be performed by referring to Table 1. For example, the value of  $\epsilon$  used in the research study Gertrude considered participating in was 0.01, and the baseline privacy risk in her opt-out scenario was 2%. As shown in Table 1, these values correspond to a worst-case privacy risk of 2.02% in her real-world scenario. Notice also how the calculation of risk would change with different values. For example, if the privacy risk in Gertrude's opt-

out scenario were 5% rather than 2% and the value of  $\epsilon$  remained the same, then the worst-case privacy risk in her real-world scenario would be 5.05%.

**\*243 Table 1. Maximal Difference Between Posterior Beliefs in Gertrude's Opt-Out and Real-World Scenarios**

The notation  $A(x')$  refers to the application of the analysis  $A$  on the dataset  $x'$ , which does not include Gertrude's information. As this table shows, the use of differential privacy provides a quantitative bound on how much one can learn about an individual from a computation.<sup>96</sup>

POSTERIOR BELIEF GIVEN $A(x')$ IN %	VALUE OF $\epsilon$					
	0.01	0.05	0.1	0.2	0.5	1
0	0	0	0	0	0	0
1	1.01	1.05	1.1	1.22	1.64	2.67
2	2.02	2.1	2.21	2.43	3.26	5.26
5	5.05	5.24	5.5	6.04	7.98	12.52
10	10.09	10.46	10.94	11.95	15.48	23.2
25	25.19	25.95	26.92	28.93	35.47	47.54
50	50.25	51.25	52.5	54.98	62.25	73.11
75	75.19	75.93	76.83	78.56	83.18	89.08
90	90.09	90.44	90.86	91.66	93.69	96.07
95	95.05	95.23	95.45	95.87	96.91	98.1
98	98.02	98.1	98.19	98.36	98.78	99.25
99	99.01	99.05	99.09	99.18	99.39	99.63
100	100	100	100	100	100	100
	maximum posterior belief given $A(x)$ in %					

The fact that the differential privacy guarantee applies to every privacy risk means that Gertrude can know for certain how participating in the study might increase her risks relative to opting out, even if she does not know a priori all the privacy risks posed by the data release. This enables Gertrude to make a more informed decision about whether to take part in the study. For instance, perhaps with the help of the researcher obtaining her informed consent, Gertrude can use this framework to better understand how the additional risk she may incur by participating in the study is bounded. By considering the bound with respect to a range of possible baseline risk values, she may **\*244** decide whether she is comfortable with taking on the risks entailed by these different scenarios.

Table 1 demonstrates how significant changes in posterior belief compared to the opt-out baseline can be for different values of  $\epsilon$ . Notice how, at  $\epsilon = 1$ , a belief that Gertrude has a certain condition with 1% probability in the opt-out scenario would become 2.67%, which is quite a large factor increase (more than double), and a 50% belief would become nearly a 75% belief (also a very significant change). For  $\epsilon = 0.2$  and  $\epsilon = 0.5$ , the changes start to become more modest, but could still be considered too large, depending on how sensitive the data are. For  $\epsilon = 0.1$  and below, the changes in beliefs may be deemed small enough for most applications.

Also note that the entries in Table 1 are the worst-case bounds that are guaranteed by a given setting of  $\epsilon$ . An adversary's actual posterior beliefs given  $A(x)$  may be smaller in a given practical application, depending on the distribution of the data, the specific differentially private algorithms used, and the adversary's prior beliefs and auxiliary information. That is, in a real-world application, a particular choice of  $\epsilon$  may turn out to be safer than Table 1 indicates, but it can be difficult to quantify how much safer.

The exact choice of  $\epsilon$  is a policy decision that should depend on the sensitivity of the data, with whom the output will be shared, the intended data analysts' accuracy requirements, and other technical and normative factors. Table 1 and explanations interpreting it, such as the examples provided in this Section, can help provide the kind of information needed to make such a policy decision.

### *E. Composition*

Privacy risk accumulates with multiple analyses on an individual's data, and this is true whether or not any privacy-preserving technique is applied.<sup>97</sup> One of the most powerful features of differential privacy is its robustness under composition.<sup>98</sup> One can reason about—and bound—the privacy risk that accumulates when multiple differentially private computations are performed on an individual's data.<sup>99</sup>

**\*245** The parameter  $\epsilon$  quantifies how privacy risk accumulates across multiple differentially private analyses. Imagine that two differentially private computations are performed on datasets about the same individuals. If the first computation uses a parameter of  $\epsilon_1$  and the second uses a parameter of  $\epsilon_2$ , then the cumulative privacy risk resulting from these computations is no greater than the risk associated with an aggregate parameter of  $\epsilon_1 + \epsilon_2$ .<sup>100</sup> In other words, the privacy risk from running the two analyses is bounded by the privacy risk from running a single differentially private analysis with a parameter of  $\epsilon_1 + \epsilon_2$ .

### *Example 14*

Suppose that Gertrude decides to opt into the medical study because it is about heart disease, an area of research she considers critically important. The study leads to a published research paper, which includes results from the study produced by a differentially private analysis with a parameter of  $\epsilon = 0.01$ . A few months later, the researchers decide that they want to use the same study data for another paper. This second paper would explore a hypothesis about acid reflux disease, and would require calculating new statistics based on the original study data. Like the analysis results in the first paper, these statistics would be computed using differential privacy, but this time with a parameter of  $\epsilon = 0.02$ .

Because she only consented to her data being used in research about heart disease, the researchers must obtain Gertrude's permission to reuse her data for the paper on acid reflux disease. Gertrude is concerned that her insurance company could compare the results from both papers and learn something negative about Gertrude's life expectancy and drastically raise her insurance premium. She is not particularly interested in participating in a research study about acid reflux disease and is concerned the risks of participation might outweigh the benefits to her.

Because the statistics from each study are produced using differentially private analyses, Gertrude can precisely bound the privacy risk that would result from contributing her data to the second study. The combined analyses can be thought of as a single analysis with a privacy loss parameter of

$$*246 \quad \epsilon_1 + \epsilon_2 = 0.01 + 0.02 = 0.03.$$

Say that, without her participation in either study, the insurance company would believe that Gertrude has a 2% chance of dying in the next year, leading to a premium of \$2,000. If Gertrude participates in both studies, the insurance company's estimate of Gertrude's mortality would increase to at most

$$2\% \cdot (1 + 0.03) = 2.06\%$$

This corresponds to a premium increase of \$60 over the premium that Gertrude would pay if she had not participated in either study.

This means that, while it cannot get around the fundamental law that privacy risk increases when multiple analyses are performed on the same individual's data, differential privacy guarantees that privacy risk accumulates in a bounded way.<sup>101</sup> Despite the accumulation of risk, two differentially private analyses cannot be combined in a way that leads to a privacy breach that is disproportionate to the privacy risk associated with each analysis in isolation. To the Authors' knowledge, differential privacy is currently the only known framework with quantifiable guarantees with respect to how risk accumulates across multiple analyses.

## V. What Types of Analyses Are Performed with differential Privacy?

A large number of analyses can be performed with differential privacy guarantees. Differentially private algorithms are known to exist for a wide range of statistical analyses such as count queries, histograms, cumulative distribution functions, and linear regression; techniques used in statistics and machine learning such as clustering and classification; and statistical disclosure limitation techniques like synthetic data generation, among many others.

For the purposes of illustrating that broad classes of analyses can be performed using differential privacy, the discussion in this Part provides a brief overview of each of these types of analyses and how they can be performed with differential privacy guarantees.<sup>102</sup>

**\*247 • Count queries:** The most basic statistical tool, a count query, returns an estimate of the number of individual records in the data satisfying a specific predicate.<sup>103</sup> For example, a count query could be used to return the number of records corresponding to HIV-positive individuals in a sample. Differentially private answers to count queries can be obtained through the addition of random noise, as demonstrated in the detailed example found in Appendix A.1.

**• Histograms:** A histogram contains the counts of data points as they are classified into disjoint categories.<sup>104</sup> For example, in the case of numerical data, a histogram shows how data are classified within a series of consecutive non-overlapping intervals. A **contingency table (or cross tabulation)** is a special form of histogram representing the interrelation between two or more variables.<sup>105</sup> The categories of a contingency table are defined as conjunctions of attribute variables, such as the number of individuals in a dataset that are both college-educated *and* earn less than \$50,000 per year.<sup>106</sup> Differentially private histograms and contingency tables provide noisy counts for the data classified in each category.<sup>107</sup>

• **Cumulative distribution function (CDF):** For data over an ordered domain, such as age (where the domain is integers, say, in the range of 0, 1, 2, ..., 100), or annual income (where the domain is real numbers, say, in the range of \$0.00 - \$1,000,000.00), a cumulative distribution function depicts for every domain value  $x$  an estimate of the number of data points with a value up to  $x$ .<sup>108</sup> A CDF can be used for computing the median of the data points \*248 (the value  $x$  for which half the data points have value up to  $x$ ) and the interquartile range, among other statistics.<sup>109</sup> A differentially private estimate of the CDF introduces noise that needs to be taken into account when the median or interquartile range is computed from the estimated CDF.<sup>110</sup>

• **Linear regression:** Social scientists are often interested in modeling how a dependent variable varies as a function of one or more explanatory variables. For instance, a researcher may seek to understand how a person's health depends on her education and income. In linear regression, an underlying linear model is assumed, and the goal of the computation is to fit a linear model to the data that minimizes a measure of "risk" (or "cost"), usually the sum of squared errors.<sup>111</sup> Using linear regression, social scientists can learn to what extent a linear model explains their data, and which of the explanatory variables correlates best with the dependent variable.<sup>112</sup> Differentially private implementations of linear regression introduce noise in its computation.<sup>113</sup>

• **Clustering:** Clustering is a data analysis technique that involves grouping data points into clusters, so that points in the same cluster are more similar to each other than to points in other clusters.<sup>114</sup> Data scientists often use clustering as an exploratory tool to gain insight into their data and identify the data's important subclasses.<sup>115</sup> Researchers are developing a variety of differentially private clustering algorithms,<sup>116</sup> and such tools are likely \*249 to be included in future privacy-preserving tool kits for social scientists.

• **Classification:** In machine learning and statistics, classification is the problem of identifying or predicting which of a set of categories a data point belongs in, based on a training set of examples for which category membership is known.<sup>117</sup> Data scientists often utilize data samples that are pre-classified (e.g., by experts or from historical data) to train a classifier, which can later be used for labeling newly acquired data samples.<sup>118</sup> Theoretical work has shown that it is possible to construct differentially private classification algorithms for a large collection of classification tasks.<sup>119</sup>

• **Synthetic data:** Synthetic data are data sets generated from a statistical model estimated using the original data.<sup>120</sup> The records in a synthetic data set have no one-to-one correspondence with the individuals in the original data set, yet the synthetic data can retain many of the statistical properties of the original data. Synthetic data resemble the original sensitive data in format, and, for a large class of analyses, results are similar whether performed on the synthetic or original data.<sup>121</sup> Theoretical work has shown that differentially private synthetic data can be generated for a large variety of tasks.<sup>122</sup> A significant benefit is that, once a differentially private synthetic data set is generated, it can be analyzed any number of times, without any further implications for



privacy.<sup>123</sup> As a result, synthetic data can be shared freely \*250 or even made public in many cases.<sup>124</sup> For example, statistical agencies can release synthetic microdata as public-use data files in place of raw microdata.<sup>125</sup>

## VI. Practical Considerations When Using Differential Privacy

This Part discusses some of the practical challenges to using differentially private computations such as those outlined in the previous Part. When making a decision regarding whether to implement differential privacy, one must consider the relevant privacy and utility requirements associated with the specific use case in mind. This Article provides many examples illustrating scenarios in which differentially private computations could be used. However, if, for instance, an analysis is being performed at the individual-level--e.g., in order to identify individual patients who would be good candidates for a clinical trial or to identify instances of bank fraud-- differential privacy would not apply, as it will disallow learning information specific to an individual.

Additionally, because implementation and use of differential privacy is in its early stages, there is a current lack of easy-to-use general purpose and production-ready tools, though progress is being made on this front, as Part VII discusses below. The literature identifies a number of other practical limitations, emphasizing the need for additional differentially private tools tailored to specific applications such as the data products released by federal statistical agencies; subject matter experts trained in the practice of differential privacy; tools for communicating the features of differential privacy to the general public, users, and other stakeholders; and guidance on setting the privacy loss parameter  $\epsilon$ .<sup>126</sup>

This Part focuses on a selection of practical considerations, including (A) challenges due to the degradation of privacy that results from composition, (B) challenges related to the accuracy of differentially private statistics, and (C) challenges related to analyzing and sharing personal data while protecting privacy in accordance with applicable \*251 regulations and policies for privacy protection. It is important to note that the challenges of producing accurate statistics, while protecting privacy and addressing composition, are not unique to differential privacy.<sup>127</sup> It is a fundamental law of information that privacy risk grows with the repeated use of data, and hence this risk applies to any disclosure limitation technique.<sup>128</sup> Traditional SDL techniques--such as suppression, aggregation, and generalization--often reduce accuracy and are vulnerable to loss in privacy due to composition.<sup>129</sup> The impression that these techniques do not suffer accumulated degradation in privacy is merely due to the fact that these techniques have not been analyzed with the high degree of rigor that differential privacy has been.<sup>130</sup> A rigorous analysis of the effect of composition is important for establishing a robust and realistic understanding of how multiple statistical computations affect privacy.<sup>131</sup>

### A. The "Privacy Budget"

As Section IV.B explains, one can think of the parameter  $\epsilon$  as determining the overall privacy protection provided by a differentially private analysis. Intuitively,  $\epsilon$  determines "how much" of an individual's privacy an analysis may utilize, or, alternatively, by how much the risk to an individual's privacy can increase. A smaller value for  $\epsilon$  implies better protection (i.e., less risk to privacy).<sup>132</sup> Conversely, a larger value for  $\epsilon$  implies worse protection (i.e., higher potential risk to privacy).<sup>133</sup> In particular,  $\epsilon = 0$  implies perfect privacy (i.e., the analysis does not increase any individual's privacy risk at all).<sup>134</sup> Unfortunately, analyses that satisfy differential privacy with  $\epsilon = 0$  must completely ignore their input data and therefore are useless.<sup>135</sup>

Section IV.B also explains that the choice of  $\epsilon$  is dependent on various normative and technical considerations, and best practices are \*252 likely to emerge over time as practitioners gain experience from working with real-world implementations of differential privacy. As a starting point, experts have suggested that  $\epsilon$  be thought of as a small value ranging from approximately 0.01 to 1.<sup>136</sup> Based on the analysis following Table 1, the Authors of this Article believe that adopting a global value of  $\epsilon = 0.1$ ,

when feasible, provides sufficient protection. In general, setting  $\epsilon$  involves making a compromise between privacy protection and accuracy. The consideration of both utility and privacy is challenging in practice and, in some of the early implementations of differential privacy, has led to choosing a higher value for  $\epsilon$ .<sup>137</sup> As the accuracy of differentially private analyses improves over time, it is likely that lower values of  $\epsilon$  will be chosen.

The privacy loss parameter  $\epsilon$  can be thought of as a “privacy budget” to be spent by different analyses of individuals' data. If a single analysis is expected to be performed on a given set of data, then one might allow this analysis to exhaust the entire privacy budget  $\epsilon$ . However, a more typical scenario is that several analyses are expected to be run on a dataset, and, therefore, one needs to calculate the total utilization of the privacy budget by these analyses.<sup>138</sup>

Fortunately, as Section IV.E discusses, a number of composition theorems have been developed for differential privacy. In particular, these theorems state that the composition of two differentially private analyses results in a privacy loss that is bounded by the sum of the privacy losses of each of the analyses.<sup>139</sup>

To understand how overall privacy loss is accounted for in this framework, consider the following example.

### *Example 15*

Suppose a data analyst using a differentially private analysis tool is required to do so while maintaining differential privacy with an overall privacy loss parameter  $\epsilon = 0.1$ . This requirement for the overall privacy loss parameter may be guided by an interpretation of a regulatory standard, institutional policy, or best practice, among other possibilities. It means that all of the analyst's analyses, taken together, must have a value of  $\epsilon$  that is at most 0.1. \*253 Consider how this requirement would play out within the following scenarios:

**One-query scenario:** The data analyst performs a differentially private analysis with a privacy loss parameter  $\epsilon_1 = 0.1$ . In this case, the analyst would not be able to perform a second analysis over the data without risking a breach of the policy limiting the overall privacy loss to  $\epsilon > 0.1$ .

**Multiple-query scenario:** The data analyst first performs a differentially private analysis with  $\epsilon_1 = 0.01$ , which falls below the limit of  $\epsilon > 0.1$ . This means that the analyst can also apply a second differentially private analysis, say with  $\epsilon_2 = 0.02$ . After the second analysis, the overall privacy loss amounts to

$$\epsilon_1 + \epsilon_2 = 0.01 + 0.02 = 0.03,$$

which is still less than  $\epsilon = 0.1$ , and therefore allows the analyst to perform additional analyses before exhausting the budget.

The multiple-query scenario can be thought of as if the data analyst has a privacy budget of  $\epsilon = 0.1$  that is consumed incrementally as she performs differentially private analyses, until the budget has been exhausted.<sup>140</sup> Performing additional analyses after the overall budget has been exhausted may result in a privacy parameter that is larger (i.e., worse) than  $\epsilon$ .<sup>141</sup> Any data use exceeding the privacy budget would result in a privacy risk that is too significant.

Note that, in the sample calculation for the multiple-query example, the accumulated privacy risk was bounded simply by adding the privacy parameters of each analysis. It is in fact possible to obtain better bounds on the accumulation of the privacy loss parameter than suggested by this example.<sup>142</sup> Various tools for calculating the bounds on the accumulated privacy risks in real-world settings using more sophisticated approaches are currently under development.<sup>143</sup>

**\*254 B. Accuracy**

This Section discusses the relationship between differential privacy and accuracy. The accuracy of an analysis is a measure of how its outcome can deviate from the true quantity or model it attempts to estimate.<sup>144</sup> There is no single measure of accuracy, as measures of deviations differ across applications.<sup>145</sup> Multiple factors have an effect on the accuracy of an estimate, including measurement and sampling errors.<sup>146</sup> The random noise introduced in differentially private computations similarly affects accuracy.<sup>147</sup>

For most statistical analyses, the inaccuracy coming from sampling error decreases as the number of samples grows,<sup>148</sup> and the same is true for the inaccuracy coming from the random noise in most differentially private analyses. In fact, it is often the case that the inaccuracy due to the random noise vanishes more quickly than the sampling error.<sup>149</sup> This means that, in theory, for very large datasets (with records for very many individuals), differential privacy comes essentially “for free.”

However, for datasets of the sizes that occur in practice, the amount of noise that is introduced for differentially private analyses can have a noticeable impact on accuracy. For small datasets, for very high levels of privacy protection (i.e., small  $\epsilon$ ), or for complex analyses, the noise introduced for differential privacy can severely impact utility.<sup>150</sup> In general, almost no utility can be obtained from datasets containing  $1/\epsilon$  or fewer records.<sup>151</sup> As Section VI.A discusses, this is **\*255** exacerbated by the fact that the privacy budget usually needs to be partitioned among many different queries or analyses, and thus the value of  $\epsilon$  used for each query needs to be much smaller. Much of the ongoing research on differential privacy is focused on understanding and improving the tradeoff between privacy and utility (i.e., obtaining the maximum possible utility from data while preserving differential privacy).<sup>152</sup>

Procedures for estimating the accuracy of certain types of analyses have been developed.<sup>153</sup> These procedures take as input the number of records, a value for  $\epsilon$ , and the ranges of numerical and categorical fields, among other parameters, and produce guaranteed accuracy bounds.<sup>154</sup> Alternatively, a desired accuracy may be given as input instead of  $\epsilon$ , and the computation results in a value for  $\epsilon$  that would provide this level of accuracy.<sup>155</sup> Figures 4(a)-(d) illustrate an example of a cumulative distribution function and the results of its noisy approximation with different settings of the privacy parameter  $\epsilon$ .<sup>156</sup>

**Figure 4. Example of the Differentially Private Computation Output**

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

**\*256** Figure 4 illustrates the outcome of a differentially private computation of the CDF of income in fictional District Q. Graph (a) presents the original CDF (without noise) and the subsequent graphs show the result of applying differentially private computations of the CDF with  $\epsilon$  values of (b) 0.005, (c) 0.01, and (d) 0.1. Notice that, as smaller values of  $\epsilon$  imply better privacy protection, they also imply less accuracy due to noise addition compared to larger values of  $\epsilon$ .

Another concept related to accuracy is truthfulness. This term has appeared regularly, if infrequently, in the statistical disclosure limitation literature since the mid-1970s, though it does not have a **\*257** well-recognized formal definition.<sup>157</sup> Roughly speaking, the SDL literature recognizes a privacy-protecting method as truthful if one can determine unambiguously which types of statements, when semantically correct as applied to the protected data (i.e., data transformed by a privacy technique such as k-anonymity), are also semantically correct when applied to the original sample data.<sup>158</sup>

This concept has an intuitive appeal. For data protected via suppressing some of the cells in the database, statements of the form “there are records with characteristics X and Y” are correct in the original data if they are correct in the protected data.

For example, one might definitively state, using only the protected data, that “some plumbers earn over \$50,000.” One cannot make this same statement definitively for data that have been synthetically generated.<sup>159</sup>

One must be careful, however, to identify and communicate the types of true statements a protection method supports. For instance, neither suppression nor synthetic data support truthful nonexistence claims at the microdata level. Even if all Wisconsin residents are included in the data, a statement such as “there are no plumbers in the dataset who earn over \$50,000” cannot be made definitively by examining the protected data alone if income or occupation values have been suppressed or synthetically generated. Moreover, protection methods may, in general, preserve truth at the individual record level, but not at the aggregate level (or vice versa).<sup>160</sup> For instance, local \*258 recoding and suppression, global recoding, and privacy criteria such as k-anonymity that use these operations in their implementation cannot produce reliably truthful statements about most aggregate computations. As an example, statements such as “the median income of a plumber in Wisconsin is \$45,000” or “the correlation between income and education in Wisconsin is .50” will not be correct.<sup>161</sup>

Assessing the truthfulness of modern privacy protection methods requires generalizing notions of truthfulness to apply to statements about the population from which the sample is drawn. Scientific research and the field of statistics are primarily concerned with making correct statements about the population.<sup>162</sup> Statistical estimates inherently involve uncertainty and, as mentioned above, there are many individual sources of error that contribute to the total uncertainty in a calculation. These are traditionally grouped by statisticians into the categories of sampling and nonsampling errors.<sup>163</sup> Correct assertions about a statistical statement accurately communicate the uncertainty of the estimated value.<sup>164</sup>

Thus, a statement is statistically truthful of protected data if it accurately communicates the uncertainty--inclusive of sampling and nonsampling errors--of the estimated population value. Methods such as local suppression and global recoding are not always capable of producing statistically truthful statements.<sup>165</sup> Fortunately, privacy \*259 protecting methods such as synthetic data generation, record swapping, and differential privacy are capable of producing statements about statistical estimates that are truthful.<sup>166</sup> For example, all of these methods could produce truthful statements such as “with a confidence level of 99%, the median income of a plumber is \$45,000 ± \$2,000.”<sup>167</sup> When produced by a truthful method, this statement correctly communicates the uncertainty of the statement, and would, roughly speaking,<sup>168</sup> turn out to be true of the population in 99 out of 100 independent trials.

Generally, differentially private methods introduce uncertainty. However, it is a property of differential privacy that the method itself does not need to be kept secret. This means the amount of noise added to the computation can be taken into account in the measure of accuracy and, therefore, lead to correct statements about the population of interest. This can be contrasted with many traditional SDL techniques, which only report sampling error and keep the information needed to estimate the “privacy error” secret. Any privacy-preserving method, if misused or misinterpreted, can produce incorrect statements. Additionally, the truthfulness of some methods, such as suppression and synthetic data generation, is inherently limited to particular levels of computations (e.g., to existence statements on microdata, or statements about selected aggregate statistical properties, respectively). Differential privacy may be used truthfully for a broader set of computations, so long as the uncertainty of each calculation is estimated and reported.

### ***C. Complying with Legal Requirements for Privacy Protection***

Statistical agencies, companies, researchers, and others who collect, process, analyze, store, or share data about individuals must take steps to protect the privacy of the data subjects in accordance with various laws, institutional policies, contracts, ethical codes, and best \*260 practices.<sup>169</sup> In some settings, tools that satisfy differential privacy can be used to analyze and share data, while both complying with legal obligations and providing strong mathematical guarantees of privacy protection for the individuals in the data.<sup>170</sup>

Privacy regulations and related guidance do not directly answer the question of whether the use of differentially private tools is sufficient to satisfy existing regulatory requirements for protecting privacy when sharing statistics based on personal data.<sup>171</sup> This issue is complex because privacy laws are often context dependent, and there are significant gaps between differential privacy and the concepts underlying regulatory approaches to privacy protection.<sup>172</sup> Different regulatory requirements are applicable depending on the jurisdiction, sector, actors, and types of information involved.<sup>173</sup> As a result, datasets held by an organization may be subject to different requirements. In some cases, similar or even identical datasets may be subject to different requirements when held by different organizations.<sup>174</sup> In addition, many legal standards for privacy protection are, to a large extent, open to interpretation and therefore require a case-specific legal analysis by an attorney.<sup>175</sup>

Other challenges arise as a result of differences between the concepts appearing in privacy regulations and those underlying differential privacy. For instance, many laws focus on the presence of “personally identifiable information” or the ability to “identify” an individual's personal information in a release of records.<sup>176</sup> Such concepts do not have precise definitions,<sup>177</sup> and their meaning in the context of differential privacy applications is especially unclear.<sup>178</sup> In addition, many privacy regulations emphasize particular requirements for protecting privacy when disclosing individual-level data, such as removing personally identifiable information, which are arguably difficult to interpret and apply when releasing aggregate statistics.<sup>179</sup> While in some cases it may be clear whether a regulatory standard has been met by the use of differential privacy, in other cases--particularly \*261 along the boundaries of a standard--there may be considerable uncertainty.<sup>180</sup> Regulatory requirements relevant to issues of privacy in computation rely on an understanding of a range of different concepts, such as personally identifiable information, de-identification, linkage, inference, risk, consent, opt out, and purpose and access restrictions. The following discussion explains how the definition of differential privacy can be interpreted to address each of these concepts while accommodating differences in how these concepts are defined across various legal and institutional contexts.

Personally identifiable information (PII) and de-identification are central concepts in information privacy law.<sup>181</sup> Regulatory protections typically extend only to personally identifiable information; information not considered personally identifiable is not protected.<sup>182</sup> Although definitions of personally identifiable information vary, they are generally understood to refer to the presence of pieces of information that are linkable to the identity of an individual or to an individual's personal attributes.<sup>183</sup> PII is also related to the concept of de-identification, which refers to a collection of techniques devised for transforming identifiable information into non-identifiable information while also preserving some utility of the data. In principle, it is intended that de-identification, if performed successfully, can be used as a tool for removing PII, or transforming PII into non-PII.<sup>184</sup>

When differential privacy is used, it can be understood as ensuring that using an individual's data will not reveal essentially any personally identifiable information specific to her.<sup>185</sup> Here, the use of the term “specific” refers to information that is unique to the individual \*262 and cannot be inferred unless the individual's information is used in the analysis.

Linkage is a mode of privacy loss recognized, implicitly or explicitly, by a number of privacy regulations.<sup>186</sup> As illustrated in Example 1, linkage typically refers to the matching of information in a database to a specific individual, often by leveraging information from external sources.<sup>187</sup> Linkage is also closely related to the concept of identifying an individual in a data release, as identifying an individual is often accomplished via a successful linkage.<sup>188</sup> Linkage has a concrete meaning when data are published as a collection of individual-level records, often referred to as microdata.<sup>189</sup> However, what is considered a successful linkage when a publication is made in other formats, such as statistical models or synthetic data, has not been defined and is open to interpretation.

Despite this ambiguity, it can be argued that differential privacy addresses record linkage in the following sense. Differentially private statistics provably hide the influence of every individual, and even small groups of individuals.<sup>190</sup> Although linkage

has not been precisely defined, linkage attacks seem to inherently result in revealing that specific individuals participated in an analysis. Because differential privacy protects against learning whether or not an individual participated in an analysis, it can therefore be understood to protect against linkage. Furthermore, differential privacy provides a robust guarantee of privacy protection that is independent of the auxiliary information available to an attacker.<sup>191</sup> Indeed, under differential privacy, even an attacker utilizing arbitrary auxiliary information cannot learn much more about an individual in a database than she could if that individual's information were not in the database at all.<sup>192</sup>

**\*263** Inference is another mode of privacy loss that is implicitly or explicitly referenced by some privacy regulations and related guidance. For example, some laws protect information that enables the identity of an individual to be “reasonably inferred,”<sup>193</sup> and others protect information that enables one to determine an attribute about an individual with “reasonable certainty.”<sup>194</sup> When discussing inference as a mode of privacy loss, it is important to distinguish between two types--inferences about individuals and inferences about large groups of individuals. Although privacy regulations and related guidance generally do not draw a clear distinction between these two types of inference,<sup>195</sup> the distinction is key to understanding which privacy safeguards would be appropriate in a given setting.

Differential privacy can be understood as essentially protecting an individual from inferences about attributes that are specific to her--that is, information that is unique to the individual and cannot be inferred unless the individual's information is used in the analysis. Interventions other than differential privacy may be necessary in contexts in which inferences about large groups of individuals, such as uses of data that result in discriminatory outcomes by race or sex, are a concern.<sup>196</sup>

Risk is another concept that appears in various ways throughout regulatory standards for privacy protection and related guidance. For example, some regulatory standards include a threshold level of risk that an individual's information may be identified in a data release.<sup>197</sup> Similarly, some regulations also acknowledge, implicitly or explicitly, that any disclosure of information carries privacy risks, and therefore the goal is to minimize, rather than eliminate, such risks.<sup>198</sup>

**\*264** Differential privacy can readily be understood in terms of risk.<sup>199</sup> Specifically, differential privacy enables a formal quantification of risk.<sup>200</sup> It guarantees that the risk to an individual is essentially the same with or without her participation in the dataset,<sup>201</sup> and this is likely true for most notions of risk adopted by regulatory standards or institutional policies. In this sense, differential privacy can be interpreted as essentially guaranteeing that the risk to an individual is minimal or very small. Moreover, the privacy loss parameter  $\epsilon$  can be tuned according to different requirements for minimizing risk.<sup>202</sup>

Consent and opt out are concepts underlying common provisions set forth in information privacy laws.<sup>203</sup> Consent and opt-out provisions enable individuals to choose to allow, or not to allow, their information to be used by or redisclosed to a third party.<sup>204</sup> Such provisions are premised on the assumption that providing individuals with an opportunity to opt in or out gives them control over the use of their personal information and effectively protects their privacy.<sup>205</sup> However, this assumption warrants a closer look. Providing consent or opt-out mechanisms as a means of providing individuals with greater control over their information is an incomplete solution as long as individuals are not fully informed about the consequences of uses or disclosures of their information.<sup>206</sup> In addition, allowing individuals the choice to opt in or out can create new privacy concerns. For example, an individual's decision to opt out may--often unintentionally--be reflected in a data release or analysis and invite scrutiny into whether the choice to opt out was motivated by the need to hide compromising information.<sup>207</sup>

The differential privacy guarantee can arguably be interpreted as providing stronger privacy protection than a consent or opt-out mechanism. This is because differential privacy can be understood as **\*265** automatically providing all individuals in the data with essentially the same protection that opting out is intended to provide.<sup>208</sup> Moreover, differential privacy provides all individuals with this privacy guarantee.<sup>209</sup> Therefore, differential privacy can be understood to prevent the possibility that

individuals who choose to opt out would, by doing so, inadvertently reveal a sensitive attribute about themselves or attract attention as individuals who are potentially hiding sensitive facts about themselves.

Purpose and access provisions often appear in privacy regulations as restrictions on the use or disclosure of personal information to specific parties or for specific purposes. Legal requirements reflecting purpose and access restrictions can be divided into two categories. The first category includes restrictions, such as those governing confidentiality for statistical agencies,<sup>210</sup> prohibiting the use of identifiable information except for statistical purposes. The second category broadly encompasses other types of purpose and access provisions, such as those permitting the use of identifiable information for legitimate educational purposes.<sup>211</sup>

Restrictions limiting use to statistical purposes, including statistical purposes involving population-level rather than individual-level analyses or statistical computations, are in many cases consistent with the use of differential privacy. This is because, as Part IV explains, differential privacy protects information specific to an individual while allowing population-level analyses to be performed. Therefore, tools that satisfy differential privacy may be understood to restrict uses to only those that are for statistical purposes, such as the definition of statistical purposes found in the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA).<sup>212</sup> However, other use and access restrictions, such as provisions limiting use to legitimate educational purposes, are orthogonal to differential privacy and require alternative privacy safeguards.<sup>213</sup>

**\*266** The foregoing interpretations of the differential privacy guarantee can be used to demonstrate that, in many cases, a differentially private mechanism would prevent the types of disclosures of personal information that privacy regulations have been designed to address. Moreover, in many cases, differentially private tools provide privacy protection that is more robust than that provided by techniques commonly used to satisfy regulatory requirements for privacy protection. However, further research to develop methods for proving that differential privacy satisfies legal requirements and setting the privacy loss parameter  $\epsilon$  based on such requirements is needed.<sup>214</sup> In practice, data providers should consult with legal counsel when considering whether differential privacy tools--potentially in combination with other tools for protecting privacy and security--are appropriate within their specific institutional settings.<sup>215</sup>

## VII. Tools for Differentially Private Analysis

At the time of this writing, differential privacy is transitioning from a purely theoretical mathematical concept to one that underlies software tools for practical use by analysts of privacy-sensitive data. The first real-world implementations of differential privacy have been deployed by companies such as Google,<sup>216</sup> Apple,<sup>217</sup> and Uber,<sup>218</sup> and government agencies such as the US Census Bureau.<sup>219</sup> Researchers in industry and academia are currently building and testing additional tools for differentially private statistical analysis. This Part briefly reviews some of these newly emerging tools, with a particular focus on the tools that inspired the drafting of this primer.

### ***\*267 A. Government and Commercial Applications of Differential Privacy***

Since 2006, the US Census Bureau has published an online interface enabling the exploration of the commuting patterns of workers across the United States, based on confidential data collected by the Bureau through the Longitudinal Employer-Household Dynamics program.<sup>220</sup> Through this interface, members of the public can interact with synthetic datasets generated from confidential survey records.<sup>221</sup> Beginning in 2008, the computations used to synthesize the data accessed through the interface have provided formal privacy guarantees that satisfy a variant of differential privacy.<sup>222</sup> In 2017, the Census Bureau announced that it was prototyping a system that would protect the full set of publication products from the 2020 decennial Census using differential privacy.<sup>223</sup>

Google, Apple, and Uber have also experimented with differentially private implementations.<sup>224</sup> For instance, Google developed the RAPPOR system, which applies differentially private computations in order to gather aggregate statistics from consumers who use the Chrome web browser.<sup>225</sup> This tool allows analysts at Google to monitor the wide-scale effects of malicious software on the browser settings of Chrome users, while providing strong privacy guarantees to individuals.<sup>226</sup> The current differentially private implementations by the Census Bureau and Uber rely on a curator model-- the model serving as the focus of most of this Article--in which a database administrator has access to and uses private data to generate differentially private data summaries.<sup>227</sup> In contrast, the current implementations by Google's RAPPOR and in Apple's macOS 10.12 and iOS 10 rely on a local model of privacy, which does not require individuals to share their private data with a trusted third party; but \*268 rather, answer questions about their own data in a differentially private manner.<sup>228</sup> Each of these differentially private answers is not useful on its own, but many of them can be aggregated to perform useful statistical analysis.

### ***B. Research and Development Towards Differentially Private Tools***

Several experimental systems from academia and industry enable data analysts to construct privacy-preserving analyses without requiring an understanding of the subtle technicalities of differential privacy. Systems such as Privacy Integrated Queries (PINQ),<sup>229</sup> Airavat,<sup>230</sup> GUPT,<sup>231</sup> Fuzz,<sup>232</sup> DFuzz,<sup>233</sup> and Ektelo<sup>234</sup> aim to provide user-friendly tools for writing programs that are guaranteed to be differentially private, through the use of differentially private building blocks<sup>235</sup> or general frameworks such as “partition-and-aggregate” or “subsample-and-aggregate”<sup>236</sup> for transforming non-private programs into differentially private ones.<sup>237</sup> These systems rely on a common approach: they keep the data safely stored and allow users to access them only via a programming interface which guarantees differential privacy.<sup>238</sup> They also afford generality, enabling one to design many types of differentially private programs that are suitable for a wide range of purposes.<sup>239</sup> However, it can be challenging for a lay user with limited expertise in programming to make effective use of these systems.<sup>240</sup>

The Authors of this Article are collaborators on the Harvard Privacy Tools Project, which develops tools to help social scientists collect, analyze, and share data while providing privacy protection for \*269 individual research subjects.<sup>241</sup> To this end, the project seeks to incorporate definitions and algorithmic tools from differential privacy into a private data-sharing interface (PSI) which facilitates data exploration and analysis using differential privacy.<sup>242</sup> PSI is intended to be integrated into research data repositories, such as Dataverse.<sup>243</sup> It will provide researchers depositing datasets into a repository with guidance on how to partition a limited privacy budget among the many statistics to be produced or analyses to be run.<sup>244</sup> It will also provide researchers seeking to explore a dataset available on the repository with guidance on how to interpret the noisy results produced by a differentially private algorithm.<sup>245</sup> Through the differentially private access enabled by PSI, researchers will be able to perform rough preliminary analyses of privacy-sensitive datasets that currently cannot be safely shared.<sup>246</sup> Such access will help researchers determine whether it is worth the effort to apply for full access to the raw data.<sup>247</sup>

### ***C. Tools for Specific Data Releases or Specific Algorithms***

There have been a number of successful applications of differential privacy with respect to specific types of data--including data from genome-wide association studies,<sup>248</sup> location history data,<sup>249</sup> data on commuter patterns,<sup>250</sup> mobility data,<sup>251</sup> client-side software data,<sup>252</sup> and data on usage patterns for phone technology.<sup>253</sup> For differentially private releases of each of these types of data, experts in differential privacy have taken care to choose algorithms and allocate privacy budgets with the aim of maximizing utility with respect to the particular data set.<sup>254</sup> Therefore, each of these tools is specific to the type of data it is designed to handle, and such tools cannot be applied in contexts in which the collection of data sources and the structure of the datasets are too heterogeneous to be compatible with such \*270 optimizations.<sup>255</sup> Thus, there remains a need for more



general-purpose tools such as those described in the previous Section. Beyond these examples, a wide literature on the design of differentially private algorithms describes approaches to performing specific data analysis tasks, including work comparing and optimizing such algorithms across a wide range of datasets. For example, the recent development of DPBench,<sup>256</sup> a framework for standardized evaluation of the accuracy of privacy algorithms, provides a way to compare different algorithms and ways of optimizing them.<sup>257</sup>

### VIII. Summary

As the previous Part illustrates, differential privacy is in initial stages of implementation in limited academic, commercial, and government settings, and research is ongoing to develop tools that can be deployed in new applications. As differential privacy is increasingly applied in practice, interest in the topic is growing among legal scholars, policymakers, and other practitioners. This Article provides an introduction to the key features of differential privacy, using illustrations that are intuitive and accessible to these audiences.

Differential privacy provides a formal, quantifiable measure of privacy. It is established by a rich and rapidly evolving theory that enables one to reason with mathematical rigor about privacy risk. Quantification of privacy is achieved by the privacy loss parameter  $\epsilon$ , which controls, simultaneously for every individual contributing to the analysis, the deviation between one's opt-out scenario and the actual execution of the differentially private analysis.

This deviation can grow as an individual participates in additional analyses, but the overall deviation can be bounded as a function of  $\epsilon$  and the number of analyses performed. This amenability to composition--or the ability to provide provable privacy guarantees with respect to the cumulative risk from successive data releases--is a unique feature of differential privacy.<sup>258</sup> While it is not the only framework that quantifies a notion of risk for a single analysis, it is currently the only framework with quantifiable guarantees on the risk resulting from a composition of several analyses.

\*271 The parameter  $\epsilon$  can be interpreted as bounding the excess risk to an individual resulting from her data being used in an analysis (compared to her risk when her data are not being used). Indirectly, the parameter  $\epsilon$  also controls the accuracy to which a differentially private computation can be performed. For example, researchers making privacy-sensitive data available through a differentially private tool may, through the interface of the tool, choose to produce a variety of differentially private summary statistics while maintaining a desired level of privacy (quantified by an accumulated privacy loss parameter), and then compute summary statistics with formal privacy guarantees.

Systems that adhere to strong formal definitions like differential privacy provide protection that is robust to a wide range of potential privacy attacks, including attacks that are unknown at the time of deployment.<sup>259</sup> An analyst designing a differentially private data release need not anticipate particular types of privacy attacks, such as the likelihood that one could link particular fields with other data sources that may be available. Differential privacy *automatically* provides a robust guarantee of privacy protection that is independent of the methods and resources used by a potential attacker.

Differentially private tools also have the benefit of transparency, as it is not necessary to maintain secrecy around a differentially private computation or its parameters. This feature distinguishes differentially private tools from traditional de-identification techniques which often require concealment of the extent to which the data have been transformed, thereby leaving data users with uncertainty regarding the accuracy of analyses on the data.

Differentially private tools can be used to provide broad, public access to data or data summaries in a privacy-preserving way. Differential privacy can help enable researchers, policymakers, and businesses to analyze and share sensitive data that cannot otherwise be shared due to privacy concerns. Further, it ensures that they can do so with a guarantee of privacy protection that substantially increases their ability to protect the individuals in the data. This, in turn, can further the progress of scientific discovery and innovation.

**Appendix A. Advanced Topics**

This Article concludes with some advanced topics for readers interested in exploring differential privacy further. This Appendix explores how differentially private analyses are constructed, explains <sup>272</sup> how the noise introduced by differential privacy compares to statistical sampling error, and discusses the protection differential privacy can provide for small groups of individuals.

**A.1. How Are Differentially Private Analyses Constructed?**

As indicated in Part IV, the construction of differentially private analyses relies on the careful introduction of uncertainty in the form of random noise. This Section provides a simple example illustrating how a carefully calibrated amount of random noise can be added to the outcome of an analysis in order to provide privacy protection.

**Example 16**

Consider computing an estimate of the number of HIV-positive individuals in a sample, where the sample contains  $n = 10,000$  individuals of whom  $m = 38$  are HIV-positive. In a differentially private version of the computation, random noise  $Y$  is introduced into the count so as to hide the contribution of a single individual. That is, the result of the computation would be  $m' = m + Y = 38 + Y$  instead of  $m = 38$ .

The magnitude of the random noise  $Y$  affects both the level of privacy protection provided and the accuracy of the count.<sup>260</sup> Generally, greater uncertainty requires a larger noise magnitude and therefore results in worse accuracy--and vice versa. In designing a release mechanism like the one described in Example 16, the magnitude of  $Y$  should depend on the privacy loss parameter  $\epsilon$ . A smaller value of  $\epsilon$  is associated with a larger noise magnitude. When choosing the noise distribution, one possibility is to sample the random noise  $Y$  from a normal distribution with zero mean and standard deviation  $1/\epsilon$ .<sup>261</sup> Because the choice of the value of  $\epsilon$  is inversely related to the magnitude of the noise introduced by the analysis, the mechanism is designed to <sup>273</sup> provide a quantifiable tradeoff between privacy and utility.<sup>262</sup> Consider the following example.

**Example 17**

A researcher uses the estimate  $m'$ , as defined in the previous example, to approximate the fraction  $p$  of HIV-positive people in the population. The computation would result in the estimate

$$p' = \frac{m'}{n} = \frac{38 + Y}{10,000}$$

For instance, suppose the sampled noise is  $Y = 4.2$ . Then, the estimate would be

$$p' = \frac{38 + Y}{10,000} = \frac{38 + 4.2}{10,000} = \frac{42.2}{10,000} = 0.42\%$$

10,000

10,000

10,000

whereas, without added noise, the estimate would have been  $p = 0.38\%$ .

**A.2 Two Sources of Error: Sampling Error and Added Noise**

This Section continues with the example from the previous Section. Note that there are two sources of error in estimating  $p$ : sampling error and added noise. The first source, sampling error, would cause  $m$  to differ from the expected  $p \cdot m$  by an amount of roughly

<<equation>>. <sup>263</sup>

For instance, consider how the researcher from the example above would calculate the sampling error associated with her estimate.

**\*274 Example 18**

The researcher reasons that  $m'$  is expected to differ from  $p \cdot 10,000$  by roughly

<<equation>>.

Hence, the estimate 0.38% is expected to differ from the true  $p$  by approximately

$$\frac{6}{10,000} = 0.06\%$$

even prior to the addition of the noise  $Y$  by the differentially private mechanism.

The second source of error is the addition of random noise  $Y$  in order to achieve differential privacy. This noise would cause  $m'$  and  $m$  to differ by an amount of roughly

$|m' - m| \approx 1/\epsilon$ . <sup>264</sup>

The researcher in the example would calculate this error as follows.

**Example 19**

The researcher reasons that, with a choice of  $\epsilon = 0.1$ , she should expect  $|m' - m| \approx 1/0.1 = 10$ , which can shift  $p'$  from the true  $p$  by

an additional  $\frac{10}{10,000} = 0.1\%$ .

10,000

Taking both sources of noise into account, the researcher calculates that the difference between noisy estimate  $p'$  and the true  $p$  is at most roughly

$$0.06\% + 0.1\% = 0.16\%$$

\*275 The two sources of noise are statistically independent,<sup>265</sup> so the researcher can use the fact that their variances add to produce a slightly better bound:

<<equation>>.

Generalizing from this example, we find that the standard deviation of the estimate  $p'$  (hence the expected difference between  $p'$  and  $p$ ) is of magnitude roughly

<<equation>>.

Notice that for a large enough sample size  $n$ , the noise added for privacy protection ( $1/n\epsilon$ ) will be much smaller than the sampling error (<<equation>>), due to the difference between having  $n$  and  $\sqrt{n}$  in the denominator, and thus privacy comes essentially “for free” in this regime. Note also that the literature on differentially private algorithms has identified many other noise introduction techniques that can result in better accuracy guarantees than the simple technique used in the examples above.<sup>266</sup> Such techniques are especially important for more complex analyses, for which the simple noise addition technique discussed in this Section is often far from optimal in terms of accuracy.

### A.3 Group Privacy

‘By holding individuals’ opt-out scenarios as the relevant baseline, the definition of differential privacy directly addresses disclosures of information localized to a single individual. However, in many cases, information may be shared between multiple individuals. For example, relatives may share an address or certain genetic attributes.

How does differential privacy protect information of this nature? Consider the opt-out scenario for a group of  $k$  individuals. This is the scenario in which the personal information of all  $k$  individuals is omitted from the input to the analysis. For instance, John and Gertrude’s opt-out scenario ( $k = 2$ ) is the scenario in which both John’s \*276 and Gertrude’s information is omitted from the input to the analysis. Recall that the parameter  $\epsilon$  controls how much the real-world scenario can differ from any individual’s opt-out scenario. It can be shown that the difference between the differentially private real-world and opt-out scenarios of a group of  $k$  individuals grows to at most

$$k \cdot \epsilon.$$
<sup>267</sup>

This means that the privacy guarantee degrades moderately as the size of the group increases. Effectively, a meaningful privacy guarantee can be provided to groups of individuals of a size of up to about

$$k \approx 1/\epsilon$$

individuals.<sup>268</sup> However, almost no protection is guaranteed to groups of

$k \approx 10/\epsilon$

individuals or greater.<sup>269</sup> This is the result of a design choice to not a priori prevent analysts using differentially private mechanisms from discovering trends across moderately-sized groups.<sup>270</sup>

#### Footnotes

<sup>a1</sup> Alexandra Wood is a Fellow at the Berkman Klein Center for Internet & Society at Harvard University. Micah Altman is Director of Research at MIT Libraries. Aaron Bembeneck is a PhD student in computer science at Harvard University. Mark Bun is a Google Research Fellow at the Simons Institute for the Theory of Computing. Marco Gaboardi is an Assistant Professor in the Computer Science and Engineering department at the State University of New York at Buffalo. James Honaker is a Research Associate at the Center for Research on Computation and Society at the Harvard John A. Paulson School of Engineering and Applied Sciences. Kobbi Nissim is a McDevitt Chair in Computer Science at Georgetown University and an Affiliate Professor at Georgetown University Law Center; work towards this document was completed in part while the Author was visiting the Center for Research on Computation and Society at Harvard University. David R. O'Brien is a Senior Researcher at the Berkman Klein Center for Internet & Society at Harvard University. Thomas Steinke is a Research Staff Member at IBM Research--Almaden. Salil Vadhan is the Vicky Joseph Professor of Computer Science and Applied Mathematics at Harvard University.

This Article is the product of a working group of the *Privacy Tools for Sharing Research Data* project at Harvard University (<http://privacytools.seas.harvard.edu>). The working group discussions were led by Kobbi Nissim. Alexandra Wood and Kobbi Nissim are the lead Authors of this Article. Working group members Micah Altman, Aaron Bembeneck, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, Salil Vadhan, and Alexandra Wood contributed to the conception of the Article and to the writing. The Authors thank John Abowd, Scott Bradner, Cynthia Dwork, Simson Garfinkel, Caper Gooden, Deborah Hurley, Rachel Kalmar, Georgios Kellaris, Daniel Muise, Michel Reymond, and Michael Washington for their many valuable comments on earlier versions of this Article. A preliminary version of this work was presented at the 9th Annual Privacy Law Scholars Conference (PLSC 2017), and the Authors thank the participants for contributing thoughtful feedback. The original manuscript was based upon work supported by the National Science Foundation under Grant No. CNS-1237235, as well as by the Alfred P. Sloan Foundation. The Authors' subsequent revisions to the manuscript were supported, in part, by the US Census Bureau under cooperative agreement no. CB16ADR0160001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the Authors and do not necessarily reflect the views of the National Science Foundation, the Alfred P. Sloan Foundation, or the US Census Bureau.

<sup>1</sup> See President's Council of Advisors on Sci. & Tech., Exec. Office of the President, *Big Data and Privacy: A Technological Perspective* (2014), [https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast\\_big\\_data\\_and\\_privacy\\_-\\_may\\_2014.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf) [<https://perma.cc/MM2V-8C2P>] (analyzing the current state of bigdata collection, storage, and use in order to make policy recommendations).

<sup>2</sup> See generally Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. Pa. L. Rev. 477 (2006) (grouping different types of privacy violations and noting their potential harms).

<sup>3</sup> See Daniel J. Solove & Paul M. Schwartz, *Information Privacy Law* 2 (6th ed. 2018).

<sup>4</sup> See *id.* at 36-38.

<sup>5</sup> See, e.g., Health Insurance Portability and Accountability Act (HIPAA), Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered titles of the U.S.C.).

<sup>6</sup> See, e.g., Family Educational Rights and Privacy Act of 1974 (FERPA), Pub. L. No. 93-380, 88 Stat. 571 (1974) (codified as amended at 20 U.S.C. § 1232g (2012)).

<sup>7</sup> See, e.g., Fair Credit Reporting Act, Pub. L. No. 91-508, 84 Stat. 1114 (1970) (codified at 15 U.S.C. §§ 1681-1681x); Gramm-Leach-Bliley Act, Pub. L. No. 106-102, 113 Stat. 1338 (1999) (codified in relevant part primarily at 15 U.S.C. §§ 6801-6809, §§ 6821-6827).

<sup>8</sup> See, e.g., Privacy Act of 1974, Pub. L. No. 93-579, 88 Stat. 1897 (1974) (codified as amended at 5 U.S.C. § 552a (2012)).

- 9 See Simson L. Garfinkel, National Institute of Standards and Technology, Deidentifying Government Datasets 46, NIST Special Publication No. 800-188 (2d Draft, 2016), [https://csrc.nist.gov/csrc/media/publications/sp/800-188/draft/documents/sp800\\_188\\_draft2.pdf](https://csrc.nist.gov/csrc/media/publications/sp/800-188/draft/documents/sp800_188_draft2.pdf) [<https://perma.cc/U6ZG-BFV5>]; Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. Rev. 1814, 1816 (2011).
- 10 See, e.g., Dep't of Health & Human Servs., Guidance Regarding Methods for Deidentification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule 6-7 (2012), [https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveridentities/Deidentification/hhs\\_deid\\_guidance.pdf](https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveridentities/Deidentification/hhs_deid_guidance.pdf) [<https://perma.cc/NRY2-M7J7>].
- 11 See Paul M. Schwartz & Edward J. Janger, *Notification of Data Security Breaches*, 105 Mich. L. Rev. 913, 972-74 (2007) (summarizing state security breach notification laws).
- 12 See Protection of Human Subjects, 45 C.F.R. §§ 46.109, .111, .116 (2018).
- 13 See, e.g., Harvard Univ. Office of the Vice Provost for Research, Harvard Research Data Security Policy (2014), [http://files.vpr.harvard.edu/files/vprdocuments/files/hrdsp\\_10\\_14\\_14\\_final\\_edits.pdf](http://files.vpr.harvard.edu/files/vprdocuments/files/hrdsp_10_14_14_final_edits.pdf) [<https://perma.cc/BDW6-T5NF>].
- 14 See Alex Kanous & Elaine Brock, Inter-Univ. Consortium for Political & Soc. Reform, Contractual Limitations on Data Sharing 3 (2015), <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/123016/ContractualLimitationsonDataSharing15041-1.pdf> [<https://perma.cc/8J AQ-LWHP>].
- 15 See, e.g., Int'l Org. for Standardization, ISO 27018 Code of Practice for Protection of Personally Identifiable Information (PII) in Public Clouds Acting as PII Processors (2014), <https://www.iso.org/standard/61498.html> [<https://perma.cc/6R3L-SH3R>] (abstract and preview).
- 16 See Commission Regulation 2016/679, 2016 O.J. (L 119) 1 [hereinafter GDPR].
- 17 See Solove & Schwartz, *supra* note 3, at 38. See generally Organisation for Economic Cooperation and Development [OECD], *Guidelines Governing the Protection of Privacy and Transborder Flow of Personal Data*, C(80)58 (July 11, 2013), <https://www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf> [<https://perma.cc/7SX3-ZEBP>] (amending 1980 version).
- 18 See, e.g., G.A. Res. 217 (III) A, Universal Declaration of Human Rights, art. 12 (Dec. 10, 1948).
- 19 See generally Fed. Comm. on Statistical Methodology, *Report on Statistical Disclosure Limitation Methodology* (Office of Mgmt. & Budget: Statistical Policy, Working Paper No. 22, 2005), <https://www.hhs.gov/sites/default/files/spwp22.pdf> [<https://perma.cc/LXN5-7QRQ>].
- 20 See *id.* at 8.
- 21 See *id.* at 12-33 (describing various SDL techniques).
- 22 See Garfinkel, *supra* note 9, at 3.
- 23 See Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701, 1716, 1731, 1742 (2010).
- 24 See *id.* at 1719-22.
- 25 See Cynthia Dwork & Aaron Roth, The Algorithmic Foundations of Differential Privacy 6-7 (2014) (originally published in 9 Found. & Trends in Theoretical Computer Sci. 211 (2014)); Garfinkel, *supra* note 9, at 47.
- 26 See *Recommendations to Identify and Combat Privacy Problems in the Commonwealth: Hearing on H.R. 351 Before the H. Select Comm. on Information Security*, 189th Sess. (Pa. 2005) [hereinafter *Pa. Privacy Hearing*] (statement of Latanya Sweeney, Associate Professor, Carnegie Mellon University), <http://dataprivacylab.org/dataprivacy/talks/Flick-05-10.html> [<https://perma.cc/W62P-Y2YX>].
- 27 See *id.*

- 28 *See id.*
- 29 *See, e.g.,* Joseph A. Calandrino et al., “You Might Also Like:” *Privacy Risks of Collaborative Filtering*, 2011 IEEE Symp. on Security & Privacy 231, 245; Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, *Nature Sci. Rep.* 4 (Mar. 25, 2013), <https://www.nature.com/articles/srep01376.pdf> [<https://perma.cc/F8DZ-347V>]; Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 IEEE Symp. on Security & Privacy 111, 123-24.
- 30 *See, e.g.,* Irit Dinur & Kobbi Nissim, *Revealing Information While Preserving Privacy*, 22 Proc. ACM SIGMOD-SIGACT-SIGART Symp. on Principles Database Sys. 202, 203-04 (2003). *See generally* Arvind Narayanan, Joanna Huey & Edward W. Felten, *A Precautionary Approach to Big Data Privacy*, in *Data Protection on the Move: Current Developments in ICT and Privacy/Data Protection* 357 (Serge Gutwirth et al. eds., 2016).
- 31 *See* Narayanan, Huey & Felten, *supra* note 30, at 366.
- 32 *See* Garfinkel, *supra* note 9, at 12, 38-40.
- 33 *See* Ohm, *supra* note 23, at 1723.
- 34 *See* Garfinkel, *supra* note 9, at 38-40.
- 35 *See* Narayanan, Huey & Felten, *supra* note 30, at 370.
- 36 *See id.* at 362-63.
- 37 As an example, in 2006, AOL published anonymized search histories of over 650,000 users over a period of three months. Shortly after the release, journalists for the *New York Times* identified a person in the release, and AOL removed the data from its web site. *See* Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, *N.Y. Times* (Aug. 9, 2006), <https://www.nytimes.com/2006/08/09/technology/09aol.html> [<https://perma.cc/GWH2-W7F8>]. However, in spite of AOL’s withdrawal of the data, copies of the data are still accessible on the internet today. *See, e.g., AOL Search Data Collection*, Internet Archive (Feb. 20, 2014), [https://archive.org/details/AOL\\_search\\_data\\_leak\\_2006](https://archive.org/details/AOL_search_data_leak_2006) [<https://perma.cc/DVX3-KPUR>].
- 38 *See generally* Cynthia Dwork et al., *Calibrating Noise to Sensitivity in Private Data Analysis*, 3 Theory Cryptography Conf. 265 (2006).
- 39 *See infra* Part V.
- 40 *See generally* Dwork et al., *supra* note 38.
- 41 *See, e.g., Differential Privacy*, Harv. U. Privacy Tools Project, <https://privacytools.seas.harvard.edu/differential-privacy> [<https://perma.cc/FA7V-NZ3K>] (last visited Sept. 14, 2018); *Putting Differential Privacy to Work*, U. Pa., <http://privacy.cis.upenn.edu> [<https://perma.cc/P5QU-XA7L>] (last visited Sept. 14, 2018).
- 42 *See* Dwork et al., *supra* note 38, at 265; *infra* Part VII.
- 43 Random noise refers to uncertainty introduced into a computation by the addition of values sampled from a random process. For example, consider a computation that first calculates the number of individuals  $x$  in the dataset who suffer from diabetes, then samples a value  $y$  from a normal distribution with a mean of 0 and variance of 1, and outputs  $z = x + y$ . In this example, the random noise  $y$  is added in the computation to the exact count  $x$  to produce the noisy output  $z$ . For a more detailed explanation of random noise, see *infra* Part IV.
- 44 *See* Dwork et al., *supra* note 38, at 266.
- 45 *See infra* Part VII.
- 46 Differential privacy was defined in 2006 by Dwork, McSherry, Nissim and Smith. Dwork et al., *supra* note 38 (building on Avrim Blum et al., *Practical Privacy: The SuLQ Framework*, 24 Proc. ACM SIGMOD-SIGACT-SIGART Symp. on Principles Database Sys. 128, 128-30 (2005); Dinur & Nissim, *supra* note 30; Cynthia Dwork & Kobbi Nissim, *Privacy-Preserving Datamining on Vertically Partitioned Databases*, 24 Ann. Int’l Cryptology Conf. 528 (2004); Alexandre Evfimievski, Johannes Gehrke,

Ramakrishnan Srikant, *Limiting Privacy Breaches in Privacy Preserving Data Mining*, 22 Proc. ACM SIGMOD-SIGACT-SIGART Symp. on Principles Database Sys. 211 (2003)). This primer's presentation of the opt-out scenario versus real-world computation is influenced by Dwork, and its risk analysis is influenced by Kasiviswanathan & Smith. Cynthia Dwork, *Differential Privacy*, 33 Int'l Colloquium on Automata, Languages & Programming 1 (2006) [hereinafter Dwork, *Differential Privacy*]; Shiva Prasad Kasiviswanathan & Adam Smith, *On the 'Semantics' of Differential Privacy: A Bayesian Formulation*, 6 J. Privacy Confidentiality 1 (2014). For other presentations of differential privacy, see Dwork (2011) and Heffetz and Ligett (2014). Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 Comm. ACM 86 (2011) [hereinafter Dwork, *A Firm Foundation*]; Ori Heffetz & Katrina Ligett, *Privacy and Data-Based Research*, 28 J. Econ. Persp. 75 (2014). For a thorough technical introduction to differential privacy, see Dwork & Roth, *supra* note 25; Salil Vadhan, *The Complexity of Differential Privacy*, in *Tutorials on the Foundations of Cryptography* 347 (Yehuda Lindell ed., 2017).

47 See Arvind Narayanan & Vitaly Shmatikov, *Myths and Fallacies of "Personally Identifiable Information"*, 53 Comm. ACM 24, 25-26 (2010). For examples illustrating what can happen if auxiliary information is not taken into account, see Narayanan, Huey & Felten, *supra* note 30, 363-65.

48 See Narayanan, Huey & Felten, *supra* note 30, at 358.

49 See *id.*; Frank McSherry, *Privacy Preserving Data Analysis*, U. Cal. Santa Cruz, [https://users.soe.ucsc.edu/~abadi/CS223\\_F12/mcsherry.pdf](https://users.soe.ucsc.edu/~abadi/CS223_F12/mcsherry.pdf) [<https://perma.cc/5DJ5-KX9B>] (last visited Oct. 4, 2018). For a general discussion of the advantages of formal privacy models over adhoc privacy techniques, see Narayanan, Huey & Felten, *supra* note 30.

50 See *Pa. Privacy Hearing*, *supra* note 26.

51 See, e.g., *supra* notes 26-29 and accompanying text.

52 See, e.g., *supra* notes 26-29 and accompanying text.

53 See, e.g., Ashwin Machanavajjhala et al., *x-Diversity: Privacy Beyond k-Anonymity*, 22 Int'l Conf. on Data Engineering 24, 24 (2006) ("In this paper we show with two simple attacks that a *k*-anonymized dataset has some subtle, but severe privacy problems.").

54 This insight follows from a series of papers demonstrating privacy breaches enabled by leakages of information resulting from decisions made by the computation. See, e.g., Krishnaram Kenthapadi, Nina Mishra & Kobbi Nissim, *Denials Leak Information: Simulatable Auditing*, 79 J. Computer & Sys. Sci. 1322, 1323 (2013).

55 One might object that the student's GPA is not traceable back to that student unless an observer knows how the statistic was produced. However, a basic principle of modern cryptography (known as Kerckhoffs' principle) holds that a system is not secure if its security depends on its inner workings being a secret. See Auguste Kerckhoffs, *La Cryptographie Militaire* [Military Cryptography] 8 (1883). As applied in this example, this means that it is taken as an assumption that the algorithm behind a statistical analysis is public (or could potentially be public).

56 See Dwork et al., *supra* note 38, at 265-66.

57 See *id.*

58 Note that these examples are introduced for the purposes of illustrating a general category of privacy-related risks relevant to this discussion, not as a claim that life insurance and mortgage companies currently engage in this practice.

59 Intuitively, the opt-out scenario and real-world scenario are very similar, and the difference between the two scenarios is measurable and small, as described in more detail in Part IV.

60 See Cynthia Dwork & Moni Naor, *On the Difficulties of Disclosure Prevention in Statistical Databases or the Case for Differential Privacy*, 2 J. Privacy & Confidentiality 93, 95 (2008).

61 See generally Dwork, *Differential Privacy*, *supra* note 46. It is important to note that the use of differentially private analysis is *not* equivalent to the traditional use of opting out. On the privacy side, differential privacy does not require an explicit opt-out. In comparison, traditional use of opt-out may cause privacy harms by calling attention to individuals who choose to opt out. On the utility side, there is no general expectation that using differential privacy would yield the same outcomes as adopting the policy of opt-out.



- 62 See Cynthia Dwork et al., *Calibrating Noise to Sensitivity in Private Data Analysis*, 7 J. Privacy & Confidentiality 17, 28 (2016) (note that this article shares a title with, and is a later version of, the authors' prior paper, *supra* note 38); Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan & Adam Smith, *Composition Attacks and Auxiliary Information in Data Privacy*, 14 Proc. ACM SIGKDD Int'l Conf. on Knowledge, Discovery & Data Mining 265, 265-66 (2008).
- 63 See *infra* Part IV.
- 64 See *infra* Part IV.
- 65 See Vadhan, *supra* note 46, at 348-49.
- 66 See *id.* at 349, 361.
- 67 See *id.*
- 68 See Fed. Comm. on Statistical Methodology, *supra* note 19, at 3.
- 69 See, e.g., Dinur & Nissim, *supra* note 30, at 203; Cynthia Dwork et al., *Exposed! A Survey of Attacks on Private Data*, 4 Ann. Rev. Stat. & Its Application 61, 64 (2016); Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, 4 PLoS Genetics e1000167, at 6, 9 (2008), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2516199/pdf/pgen.1000167.pdf> [<https://perma.cc/7873-CG6L>]; Fed. Comm. on Statistical Methodology, *supra* note 19, at 3.
- 70 See sources cited *supra* note 69.
- 71 See sources cited *supra* note 69.
- 72 See sources cited *supra* note 69.
- 73 Ellen's inference would rely on factors such as the size of the study sample, whether the sampling was performed at random, and whether John comes from the same population as the sample, among others.
- 74 See Dwork et al., *supra* note 38, at 265-66.
- 75 See *id.* at 267.
- 76 Dwork et al., *supra* note 62, at 18.
- 77 Note that, if an analyst is allowed to repeat this computation multiple times, she could average out the noise and get the exact answer. The number of allowable repetitions is limited by an overall privacy budget. See *infra* Section VI.A.
- 78 See, e.g., Dwork & Roth, *supra* note 25, at 22; Prashanth Mohan et al., *GUPT: Privacy Preserving Data Analysis Made Easy*, 2012 Proc. ACM SIGMOD Int'l Conf. on Mgmt. Data 349, 349; Vadhan, *supra* note 46, at 366-67; Marco Gaboardi et al., *PSI ( $\psi$ ): A Private Data Sharing Interface* 15 (ArXiv, Working Paper No. 1609.04340, 2018), <https://arxiv.org/pdf/1609.04340.pdf> [<https://perma.cc/PXC4-6CEL>].
- 79 See Dwork & Roth, *supra* note 25, at 6.
- 80 *Id.*
- 81 See *id.*
- 82 The property that differential privacy is preserved under arbitrary further processing is referred to as (resilience to) post-processing. See Dwork & Roth, *supra* note 25, at 19.
- 83 See *id.* For an illustration of how the choice of epsilon can affect accuracy, see *infra* Figure 4.
- 84 See *infra* Figure 4.

- 85 See John M. Abowd & Ian M. Schmutte, Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods 1 (2015), <https://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1036&context=Idi> [<https://perma.cc/8B8Q-LCFA>]; Garfinkel, *supra* note 9, at 54; Justin Hsu et al., *Differential Privacy: An Economic Method for Choosing Epsilon*, 27 IEEE Computer Security Found. Symp. 398, 398 (2014). See generally John M. Abowd & Ian M. Schmutte, *An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices*, Am. Econ. Rev. (forthcoming).
- 86 Setting the primary loss parameter  $\epsilon$  is a policy decision to be informed by normative and technical considerations. Companies and governments experimenting with practical implementations of differential privacy have selected various values for  $\epsilon$ . Some of these implementations have adopted values of  $\epsilon$  exceeding 1 due to the difficulty of meeting utility requirements using lower values of  $\epsilon$ . To date, these choices of  $\epsilon$  have not led to known vulnerabilities. For example, the US Census Bureau reportedly chose a value of  $\epsilon = 8.9$  for OnTheMap—a public interface which allows users to explore American commuting patterns using a variant of differential privacy. See John M. Abowd, Assoc. Dir. for Research and Methodology, US Census Bureau, The Challenge of Scientific Reproducibility and Privacy Protection for Statistical Agencies, Presentation for the Census Scientific Advisory Committee 12 (Sept. 15, 2016), <https://www2.census.gov/cac/sac/meetings/2016-09/2016-abowd.pdf> [<https://perma.cc/4CXN-C257>]. As another example, researchers have determined that Apple's differential private data collection in macOS 10.12 and iOS 10 likely uses values of  $\epsilon$  as high as 6 and 14, respectively. See Jun Tang et al., *Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12* (ArXiv, Working Paper No. 1709.02753, 2017), <https://arxiv.org/pdf/1709.02753.pdf> [<https://perma.cc/V4QE-QJ49>]. Although differential privacy is an emerging concept and has been deployed in limited applications to date, best practices may emerge over time as values for  $\epsilon$  are selected for implementations of differential privacy in a wide range of settings. With this in mind, researchers have proposed that a registry be created to document details of differential privacy implementations, including the value of  $\epsilon$  chosen and the factors that led to its selection. See Nat'l Acad. of Scis., Eng'g & Med., Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps 107 (Robert M. Groves & Brian A. Harris-Kojetin eds., 2017) (citing Cynthia Dwork & Dierdre Mulligan, *Differential Privacy in Practice: Expose Your Epsilons!* (June 5, 2014) (unpublished manuscript)), <http://nap.edu/24893> [<https://perma.cc/5YKH-QQBG>].
- 87 See discussion following Table 1.
- 88 See *supra* Part III.
- 89 See Dwork et al., *supra* note 38, at 266-67.
- 90 Figures in this example are based on data from *Actuarial Life Table: Period Life Table, 2015*, Soc. Security Admin., <http://www.ssa.gov/oact/STATS/table4c6.html> [<https://perma.cc/7ZPH-GE7N>] (last visited Sept. 22, 2018).
- 91 Note that there may be legal, policy, or other reasons why a company would not raise Gertrude's insurance premium based on the outcome of this study. Also, this is not a claim that insurance companies engage in this practice. Example 11 is introduced for the purposes of illustrating a general category of privacy-related risks relevant to this discussion. This example assumes that the insurance company updates its belief about Gertrude's chances of dying next year based on the outcome of this study using a Bayesian analysis. Furthermore, it assumes that Gertrude's premium is then updated in proportion to this change in belief. Differential privacy also allows one to reason (in a different manner) about a more general case where no assumptions are made regarding how the insurance company updates Gertrude's premium, but that analysis is omitted from this discussion for simplicity.
- 92 Although Gertrude, acting as an individual, cannot avoid this risk, society or groups of individuals may collectively act to avoid such a risk. For example, the researchers could be prohibited from running the study, or the data subjects could collectively decide not to participate. Therefore, the use of differential privacy does not completely eliminate the need to make policy decisions regarding the value of allowing data collection and analysis in the first place.
- 93 See Dwork et al., *supra* note 62, at 19; Dwork & Naor, *supra* note 60, at 103.
- 94 In general, the guarantee made by differential privacy is that the probabilities differ by at most a factor of  $e^{\pm\epsilon}$ , which is approximately  $1 \pm \epsilon$  when  $\epsilon$  is small. See Shiva Prasad Kasiviswanathan & Adam Smith, *On the 'Semantics' of Differential Privacy: A Bayesian Formulation*, 6 J. Privacy & Confidentiality 1 (2014).
- 95 See *infra* Table 1 and accompanying text.
- 96 For  $p$ , the posterior belief given  $A(x')$ , and privacy parameter  $\epsilon$ , the bound on the posterior

belief given  $A(x)$  is  $p$  . For small  $\epsilon$  and  $p$ , this expression can be approximated as  $p(1 + \epsilon)$ . These

$$p + e^{-\epsilon}(1-p)$$

formulas are derived from the definition of differential privacy. See Kobbi Nissim, Claudio Orlandi & Rann Smorodinsky, *Privacy-Aware Mechanism Design*, 13 Proc. ACM Conf. on Electronic Com. 774, 775-89 (2012).

97 See Dwork & Roth, *supra* note 25, at 5. Note that this observation is not unique to differentially private analyses. It is true for *any* use of information, and, therefore, for any approach to preserving privacy. However, the fact that the cumulative privacy risk from multiple analyses can be bounded is a distinguishing property of differential privacy.

98 See sources cited *supra* note 62.

99 See sources cited *supra* note 62.

100 See Dwork et al., *supra* note 62, at 28.

101 See *id.* at 28-29.

102 The discussion in this Part provides only a brief introduction to a number of statistical and machine learning concepts. For a more detailed introduction to these concepts, see, for example, Joseph K. Blitzstein & Jessica Hwang, *Introduction to Probability* (2015); Gareth James et al., *An Introduction to Statistical Learning with Applications in R* 127-75 (2013).

103 See Mark Bun, *A Teaser for Differential Privacy 1* (Dec. 8, 2017) (unpublished manuscript), <https://www.cs.princeton.edu/~smattw/Teaching/521fa17lec22.pdf> [<https://perma.cc/L54G-BKUW>].

104 See John M. Chambers et al., *Graphical Methods for Data Analysis* 24-26 (1983).

105 See Yvonne M. Bishop, Stephen E. Fienberg & Paul W. Holland, *Discrete Multivariate Analysis: Theory and Practice* 9-13 (1975).

106 See *id.*

107 See, e.g., Dwork et al., *supra* note 38, at 273.

108 See James E. Gentle, *Computational Statistics* 29-30 (2009).

109 See *id.* at 62-63, 330.

110 For a more in-depth discussion of differential privacy and CDFs, see Daniel Muise & Kobbi Nissim, Ctr. for Research on Computation & Soc'y, *Presentation at Harvard University: Differential Privacy in CDFs* (Apr. 2016), [http://privacytools.seas.harvard.edu/files/dpcdf\\_user\\_manual\\_aug\\_2016.pdf](http://privacytools.seas.harvard.edu/files/dpcdf_user_manual_aug_2016.pdf) [<https://perma.cc/DZU8-7SSB>] (slide deck).

111 See William H. Green, *Econometric Analysis* 13-14, 28-29 (8th ed. 2017).

112 See *id.*

113 See, e.g., Adam Smith, *Privacy-Preserving Statistical Estimation with Optimal Convergence Rates*, 43 Proc. ACM Symp. on Theory Computing 813, 814 (2011).

114 See Trevor Hastie, Robert Tibshirani & Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, & Prediction* 501 (2d ed. 2001).

115 See *id.* at 502.

116 Many papers describe differentially private clustering algorithms. For a recent example, see Haim Kaplan & Uri Stemmer, *Differentially Private k-Means with Constant Multiplicative Error 1* (ArXiv, Working Paper No. 1804.08001, 2018), <https://arxiv.org/abs/1804.08001> [<https://perma.cc/HR35-FHHK>].

- 117 See James et al., *supra* note 102, at 127-29.
- 118 See *id.*
- 119 Many papers describe differentially private classification algorithms. For an early example, see Blum et al., *supra* note 46.
- 120 See Jerome P. Reiter, *Satisfying Disclosure Restrictions with Synthetic Data Sets*, 18 J. Official Stat. 531, 531 (2002); Jerome P. Reiter & Trivellore E. Raghunathan, *The Multiple Adaptations of Multiple Imputation*, 102 J. Am. Stat. Ass'n 1462, 1466 (2007); Donald B. Rubin, Discussion, *Statistical Disclosure Limitation*, 9 J. Official Stat. 461, 464 (1993).
- 121 See Rubin, *supra* note 120, at 463.
- 122 See, e.g., Avrim Blum, Katrina Ligett & Aaron Roth, *A Learning Theory Approach to Non-Interactive Database Privacy*, 40 Proc. ACM Symp. on Theory Computing 609, 609 (2008).
- 123 See Nat'l Acads. of Scis., Eng'g & Med., *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy* 94 (Robert M. Groves & Brian A. Harris-Kojetin eds., 2017).
- 124 For an example of public use synthetic microdata, see Ashwin Machanavajjhala et al., *Privacy: Theory Meets Practice on the Map*, 24 Proc. IEEE Int'l Conf. on Data Engineering 277, 277 (2008).
- 125 See Ron S. Jarmin, Thomas A. Louis & Javier Miranda, *Expanding the Role of Synthetic Data at the U.S. Census Bureau* 3 (Ctr. for Econ. Studies, Research Paper No. CES 14-10, 2014), <https://www2.census.gov/ces/wp/2014/CES-WP-14-10.pdf> [<https://perma.cc/6UXHTMKM>].
- 126 See Simson L. Garfinkel, John M. Abowd & Sarah Powazek, *Issues Encountered Deploying Differential Privacy* (ArXiv, Working Paper No. 1809.02201, 2018), <https://arxiv.org/abs/1809.02201> [<https://perma.cc/4FL6-JU46>].
- 127 See Dwork et al., *supra* note 62, at 82.
- 128 See *id.*
- 129 See Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan & Adam Smith, *Composition Attacks and Auxiliary Information in Data Privacy*, 14 Proc. ACM SIGKDD Int'l Conf. on Knowledge, Discovery & Data Mining 265, 265-66 (2008).
- 130 For a discussion of privacy and utility with respect to traditional statistical disclosure limitation techniques, see generally Bee-Chung Chen et al., *Privacy-Preserving Data Publishing*, 2 Found. & Trends in Databases 1 (2009). As shown in Example 5, techniques relying on aggregation do not necessarily compose well. Furthermore, this phenomenon has been demonstrated more generally with respect to a wide range of traditional statistical disclosure limitation techniques. See generally Ganta, Kasiviswanathan & Smith, *supra* note 129.
- 131 See *id.* at 266.
- 132 See Dwork et al., *supra* note 62, at 18.
- 133 See *id.* at 18.
- 134 See *id.*
- 135 See *supra* Part IV.B.
- 136 See, e.g., Dwork, *A Firm Foundation*, *supra* note 46, at 91 (“[W]e tend to think of  $\epsilon$  as, say, 0.01, 0.1, or in some cases,  $\ln 2$  or  $\ln 3$ .”).
- 137 See *supra* notes 85-86 and the discussion following Table 1.
- 138 See Heffetz & Ligett, *supra* note 46, at 84 (discussing various examples in which the privacy budget is divided across several analyses).
- 139 See Dwork et al., *supra* note 62, at 28.

- 140 See Heffetz & Ligett, *supra* note 46, at 84.
- 141 See *id.* at 84, 87.
- 142 A number of papers explore ways to improve these bounds. See, e.g., Amos Beimel, Kobbi Nissim & Eran Omri, *Distributed Private Data Analysis: Simultaneously Solving How and What*, 2008 Advances in Cryptography (CRYPTO) 451; Cynthia Dwork, Guy N. Rothblum & Salil Vadhan, *Boosting and Differential Privacy*, 51 IEEE Ann. Symp. on Found. Computer Sci. 51 (2010); Peter Kairouz, Sewoong Oh & Pramod Viswanath, *The Composition Theorem for Differential Privacy*, 63 IEEE Transactions on Info. Theory 4037 (2017); Jack Murtagh & Salil P. Vadhan, *The Complexity of Computing the Optimal Composition of Differential Privacy*, 2016 Theory of Cryptography 157.
- 143 See Gaboardi et al., *supra* note 78, at 7.
- 144 See Int'l Statistical Inst., *The Oxford Dictionary of Statistical Terms* 4 (Yadolah Dodge ed., 6th ed. 2006).
- 145 For example, a researcher interested in estimating the average income of a given population may care about the absolute error of this estimate (i.e., the difference between the real average and the estimate), whereas a researcher interested in the median income may care about the difference between the number of respondents whose income is below the estimate and the number of respondents whose income is above the estimate.
- 146 Measurement error is the difference between the measured value of a quantity and its true value (e.g., an error in measuring an individual's height or weight), and sampling error is error caused by observing a sample rather than the entire population (e.g., the fraction of people with diabetes in the sample is likely to be different from the fraction with diabetes in the population).
- 147 See Muise & Nissim, *supra* note 110, at 94.
- 148 See Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* 6 (1977).
- 149 See generally Dwork et al., *supra* note 62; Smith, *supra* note 113; *infra* Appendix A.2.
- 150 See Muise & Nissim, *supra* note 110; Michael Hay et al., *Principled Evaluation of Differentially Private Algorithms Using DPBench*, 2016 Proc. ACM SIGMOD Int'l Conf. on Mgmt. Data 139, 139, <http://dl.acm.org/citation.cfm?id=2882931> [<https://perma.cc/6BQD-PQCT>].
- 151 This rule of thumb follows directly from the definition of differential privacy. See Dwork et al., *supra* note 62, at 17, 18. Specifically, the parameter  $\epsilon$  bounds the distance between the probability distributions resulting from a differentially private computation on two datasets that differ on one entry. Datasets containing only  $1/\epsilon$  entries can differ on at most this number of entries. Summing the differences over just  $1/\epsilon$  entries reveals that, for any two datasets of this size, the differentially private mechanism produces distributions that are at distance  $\epsilon \cdot 1/\epsilon = 1$  at most. A distance of this size would usually not support any reasonable utility.
- 152 See, e.g., Dwork, *Differential Privacy*, *supra* note 46, at 6; Dwork & Roth, *supra* note 25, at 158; Vadhan, *supra* note 46, at 58-59, 77.
- 153 See Mohan et al., *supra* note 78, at 349; Gaboardi et al., *supra* note 78, at 15.
- 154 See Gaboardi et al., *supra* note 78, at 15.
- 155 See *id.* at 12, 15.
- 156 Figures 4(a)-(d) are adapted from Muise & Nissim, *supra* note 110, at 113.
- 157 See, e.g., Lawrence H. Cox & Gordon Sande, *Techniques for Preserving Statistical Confidentiality*, 42 Proc. Int'l Stat. Inst. 6 (1979); Josep Domingo-Ferrer, David Sánchez & Jordi Soria-Comas, *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-Based Inter-Model Connections*, 15 Synthesis Lectures Info. Security, Privacy & Tr. 1, 15 (2016) (distinguishing between “perturbative masking (which distorts the original data and leads to the publication of non-truthful data) and non-perturbative masking (which reduces the amount of information, either by suppressing some of the data or by reducing the level of detail, but preserves truthfulness)”; Benjamin C. M. Fung et al., *Privacy Preserving Data Publication: A Survey of Recent Developments*, 42 ACM Computing Survs., no. 14, 2010, at 4 (describing, without defining, truthfulness at the record level by explaining that “[i]n some

data publishing scenarios, it is important that each published record corresponds to an existing individual in real life .... Randomized and synthetic data do not meet this requirement. Although an encrypted record corresponds to a real life patient, the encryption hides the semantics required for acting on the patient represented.”).

- 158 See sources cited *supra* note 157. Note that this definition of truthfulness is analogous to the general notion of avoiding false precision and is consistent with recognized principles for reporting statistical results. See, e.g., Tom Lang & Douglas Altman, *Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines*, 25 Medical Writing 31 (2016).
- 159 Synthetic data generation, by definition, uses a statistical model built from one set of data to generate new data. This preserves some of the statistical characteristics of the data, but not the original records themselves. See Fung et al., *supra* note 157, at 4. As a result, any measurement made on the synthetic dataset is related only probabilistically to measurements made on the original data and is associated with a measure of uncertainty.
- 160 See generally A. F. Karr et al., *A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality*, 60 Am. Statistician 224 (2006) (discussing various approaches to evaluating the utility of data protected by statistical disclosure limitation techniques).
- 161 Correctly calculating and truthfully reporting the uncertainty induced by suppression would require revealing the full details of the suppression algorithm and its parameterization. Revealing these details allows information to be inferred about individuals. Traditional SDL techniques require that the mechanism itself be kept secret in order to protect against this type of attack.
- 162 In general terms, the goal of statistics is to make reliable inferences about a population or distribution based on characteristics calculated from a sample of data drawn from that population. For a mathematically detailed definition, see Allan Birnbaum, *On the Foundations of Statistical Inference*, 57 J. Am. Stat. Ass'n 269, 273 (1962). In similarly general terms, the goal of science is to yield reliable generalized knowledge about the world, such as knowledge about populations, general predictions, or natural laws. A widely recognized example capturing this distinction is the regulatory definition of scientific research found in the Federal Policy for the Protection of Human Subjects. See 45 C.F.R. § 46.102(1) (2018) (“Research means a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge.”).
- 163 See *Error Measurement*, Bureau of Lab. Stat., <https://www.bls.gov/opub/hom/topic/error-measurements.htm> [<https://perma.cc/66U6-HJFA>] (last visited Sept. 13, 2018).
- 164 See Micah Altman, Jeff Gill & Michael P. McDonald, *Numerical Issues in Statistical Computing for the Social Scientist* 260-61 (2004).
- 165 See Leon Willenborg & Ton de Waal, *Elements of Statistical Disclosure Control* 28 (2001) (discussing how SDL techniques may introduce bias). For instance, Willenborg and de Waal note specifically that suppression of local values (i.e., cells, when used in the context of microdata) induces missing-data bias. Generalization takes many forms, and these forms are associated with different sources of statistical bias. For example, range generalization (e.g., top-coding) involves collapsing the observed distribution of values, which statisticians recognize as yielding truncation bias, whereas global recoding to suppress an entire measure may induce missing-variable bias in a subsequently estimated model. See generally Jack Johnston & John DiNardo, *Econometric Methods* (4th ed. 1996) (discussing these types of biases).
- 166 Each of these methods can be applied in such a way that correctly calibrated measures of uncertainty accompany computed statistics. For a detailed treatment of using differential privacy to carefully calibrate the uncertainty in statistical estimates, see Cynthia Dwork et al., *The Reusable Holdout: Preserving Validity in Adaptive Data Analysis*, 349 Sci. 636 (2015).
- 167 From this statement, we can derive other conclusions, such as that, with 99% confidence, at least half of all plumbers earn over \$43,000 annually. And if existence statements such as these are the main concern, one could use other differentially private algorithms to support making similar statements with near certainty--not merely 99% confidence.
- 168 For a precise treatment of frequentist statistical confidence intervals, see D.R. Cox & D.V. Hinkley, *Theoretical Statistics* 48-49, 208-09 (1974).
- 169 See *supra* Section I.A (discussing legal and ethical frameworks for data privacy).
- 170 See Kobbi Nissim et al., *Bridging the Gap Between Computer Science and Legal Approaches to Privacy*, 31 Harv. J.L. & Tech. 687, 697 (2018).

- 171 *See id.* at 733.
- 172 *See id.* at 730, 735.
- 173 *See id.* at 691; Schwartz & Solove, *supra* note 9, at 1847.
- 174 *See* Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 Berkeley Tech. L.J. 1967, 2009 (2015).
- 175 *See id.* at 1972.
- 176 *See* Schwartz & Solove, *supra* note 9, at 1816.
- 177 *See id.*
- 178 *See* Nissim et al., *supra* note 170, at 691, 730-31.
- 179 *See id.* at 720.
- 180 *See id.* at 710.
- 181 *See* Schwartz & Solove, *supra* note 9, at 1819.
- 182 *See id.* at 1816.
- 183 For a survey of various definitions of *personally identifiable information*, see *id.* at 1829-36. The Government Accountability Office also provides a general definition of personally identifiable information. *See* U.S. Gov't Accountability Office, GAO-08-536, Alternatives Exist for Enhancing Protection of Personally Identifiable Information (2008) (“For purposes of this report, the terms *personal information* and *personally identifiable information* are used interchangeably to refer to any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, Social Security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information.”), <https://www.gao.gov/new.items/d08536.pdf> [<https://perma.cc/9DTU-H7S6>].
- 184 *See, e.g.*, 34 C.F.R. § 99.31(b)(1) (2018) (provision for “[d]e-identified records and information,” which permits the release of education records “after the removal of all personally identifiable information provided that the educational agency or institution or other party has made a reasonable determination that a student's identity is not personally identifiable, whether through single or multiple releases, and taking into account other reasonably available information”).
- 185 Note that the reference to “using an individual's data” in this statement means the inclusion of an individual's data in an analysis.
- 186 For example, by defining personally identifiable information in terms of information “linked or linkable to a specific student,” FERPA appears to emphasize the risk of a successful record linkage attack. *See* 34 C.F.R. § 99.3 (2018). The Department of Health & Human Services in guidance on de-identifying data in accordance with the HIPAA Privacy Rule includes an extended discussion of examples of record linkage attacks and de-identification strategies for mitigating them. *See* Dep't of Health & Human Servs., *supra* note 10, at 15-17. Guidance on complying with European data protection law refers to linkability, “which is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases),” as one of three risks essential to anonymization. *Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques*, at 11 (Apr. 10, 2014) [hereinafter *Article 29 Data Protection Working Party*].
- 187 *See* Dwork & Roth, *supra* note 25, at 6-7; Fed. Comm. on Statistical Methodology, *supra* note 19, at 83.
- 188 *See* sources cited *infra* note 186.
- 189 *See* Fed. Comm. on Statistical Methodology, *supra* note 19, at 4.
- 190 *See* Dwork et al., *supra* note 62, at 17, 29.

- 191 See Ganta, Kasiviswanathan & Smith, *supra* note 129, at 265.
- 192 See *id.* at 271.
- 193 See, e.g., E-Government Act of 2002, [Pub. L. 107-347, 116 Stat. 2899](#), § 208 (2002) (codified as amended at [44 U.S.C. § 3501 \(2012\)](#)) (“[T]he term ‘identifiable form’ means any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means.”).
- 194 See, e.g., [34 C.F.R. § 99.3 \(2018\)](#) (defining “personally identifiable information,” in part, in terms of information that would allow one to identify a student “with reasonable certainty”).
- 195 See, e.g., *Article 29 Data Protection Working Party*, *supra* note 186, at 12 (defining inference broadly as “the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes”).
- 196 See Micah Altman et al., *Practical Approaches to Big Data Privacy Over Time*, 8 Int’l Data Privacy L. 29, 43 (2018); Micah Altman, Alexandra Wood & Effy Vayena, *A Harm-Reduction Framework for Algorithmic Fairness*, 16 IEEE Security & Privacy 34 (2018).
- 197 The HIPAA Privacy Rule requires covered entities to use de-identification techniques prior to releasing data in order to create a dataset with only a “very small” risk of identification. [45 C.F.R. § 164.514\(b\)\(1\) \(2018\)](#).
- 198 Guidance on complying with the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) requires agencies to “[c]ollect and handle confidential information to minimize risk of disclosure.” See [Implementation Guidance for Title V of the E-Government Act, 72 Fed. Reg. 33,362-33,363 \(June 15, 2007\)](#). Guidance from the Department of Health & Human Services recognizes that de-identification methods “even when properly applied, yield de-identified data that retains some risk of identification. Although the risk is very small, it is not zero, and there is a possibility that de-identified data could be linked back to the identity of the patient to which it corresponds.” Dep’t of Health & Human Servs., *supra* note 10, at 6.
- 199 See *supra* Section IV.C.
- 200 See *id.*
- 201 See *id.*
- 202 See *supra* Section IV.B.
- 203 See generally Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 Harv. L. Rev. 1880, 1884, 1901 (2013).
- 204 See, e.g., [34 C.F.R. § 99.37 \(2018\)](#) (including a provision requiring educational agencies and institutions to offer students an opportunity to opt out of the disclosure of their personal information in school directories).
- 205 See Solove, *supra* note 203, at 1880.
- 206 See *id.* at 1885.
- 207 See, e.g., Kim Zetter, *The NSA Is Targeting Users of Privacy Services, Leaked Code Shows*, *Wired* (July 3, 2014, 5:45 PM), <https://www.wired.com/2014/07/nsa-targets-users-of-privacy-services/> [<https://perma.cc/2KVL-LKS4>] (revealing that the National Security Agency’s surveillance efforts specially target users of privacy services).
- 208 See *supra* Part IV.
- 209 See *id.*
- 210 See, e.g., Confidential Information Protection and Statistical Efficiency Act of 2002, [Pub. L. No. 107-347, 116 Stat. 2899](#), 2963, 2966 (2002) (codified as amended at [44 U.S.C. § 3501 \(2012\)](#)) (prohibiting the use of protected information “for any use other than an exclusively statistical purpose,” where *statistical purpose* “means the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups”).



- 211 For example, FERPA generally prohibits the disclosure of personally identifiable information from education records, with limited exceptions such as disclosures to school officials with a legitimate educational interest in the information, 34 C.F.R. § 99.31(a)(1) (2018), or to organizations conducting studies for, or on behalf of, schools, school districts, or postsecondary institutions, § 99.31(a) (6).
- 212 See *supra* note 210.
- 213 See Altman et al., *supra* note 196, at 47.
- 214 For an extended discussion of the gaps between legal and computer science definitions of privacy and a demonstration that differential privacy can be used to satisfy an institution's obligations under FERPA, see Nissim et al., *supra* note 170.
- 215 For a framework for selecting among differential privacy and other suitable privacy and security controls, see Altman et al., *supra* note 196, at 29; Altman et al., *supra* note 174, at 2022.
- 216 See Úlfar Erlingsson, Vasył Pihur & Aleksandra Korolova, *RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response*, 2014 Proc. ACM Conf. on Computer & Comm. Security 1054, 1055 (2014) [hereinafter Erlingsson et al., *RAPPOR*]; Úlfar Erlingsson, *Learning Statistics with Privacy, Aided by the Flip of a Coin*, Google AI Blog (Oct. 30, 2014), <http://googleresearch.blogspot.com/2014/10/learning-statistics-with-privacy-aided.html> [https://perma.cc/Q873-TZZS].
- 217 Andy Greenberg, *Apple's 'Differential Privacy' Is About Collecting Your Data--But Not Your Data*, Wired (June 13, 2016, 7:02 PM), <http://www.wired.com/2016/06/apples-differential-privacy-collecting-data/> [https://perma.cc/5A47-GP96].
- 218 See Noah Johnson, Joseph P. Near & Dawn Song, *Towards Practical Differential Privacy for SQL Queries*, 11 Proc. VLDB Endowment 526, 526 (2018).
- 219 See *OnTheMap Application for the Longitudinal Employer-Household Dynamics Program*, US Census Bureau, <http://onthemap.ces.census.gov> [https://perma.cc/WNX3-CQFB] (last visited Sept. 25, 2018).
- 220 See *id.*
- 221 See *OnTheMap Help and Documentation*, US Census Bureau, <https://lehd.ces.census.gov/applications/help/onthemap.html#!faqs> [https://perma.cc/P7PU-4CL2] (last visited Oct. 4, 2018).
- 222 See Machanavajjhala et al., *supra* note 124, at 277.
- 223 See generally Garfinkel, Abowd & Powazek, *supra* note 126.
- 224 See Erlingsson et al., *RAPPOR*, *supra* note 216; Greenberg, *supra* note 217; Johnson, Near & Song, *supra* note 218, at 526.
- 225 See Erlingsson et al., *RAPPOR*, *supra* note 216.
- 226 *Id.* Other examples for using differential privacy (for which, to the best of the Authors' knowledge, no technical reports have been published) include Google's use of differential privacy in analyzing urban mobility and Apple's use of differential privacy in iOS 10. See Andrew Eland, *Tackling Urban Mobility with Technology*, Google Eur. Blog (Nov. 18, 2015), <http://googlepolicyeurope.blogspot.com/2015/11/tackling-urban-mobility-with-technology.html>; Greenberg, *supra* note 217.
- 227 See Garfinkel, Abowd & Powazek, *supra* note 126; Johnson, Near & Song, *supra* note 218.
- 228 See Erlingsson et al., *RAPPOR*, *supra* note 216; Greenberg, *supra* note 217.
- 229 Frank McSherry, *Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis*, 2009 Proc. ACM SIGMOD Int'l Conf. on Mgmt. Data 19, 19-20.
- 230 Indrajit Roy et al., *Airavat: Security and Privacy for MapReduce*, USENIX (2010), [http://www.usenix.org/events/nsdi10/tech/full\\_papers/roy.pdf](http://www.usenix.org/events/nsdi10/tech/full_papers/roy.pdf) [https://perma.cc/N6FF-8SSB].
- 231 Mohan et al., *supra* note 78, at 349-50.

- 232 Jason Reed & Benjamin C. Pierce, *Distance Makes the Types Grow Stronger: A Calculus for Differential Privacy*, 15 ACM SIGPLAN Int'l Conf. on Functional Programming 157 (2010).
- 233 Marco Gaboardi et al., *Linear Dependent Types for Differential Privacy*, 40 Proc. Ann. ACM SIGPLAN-SIGACT Symp on Principles Programming Languages 357 (2013).
- 234 Dan Zhang et al., *EKTELO: A Framework for Defining Differentially-Private Computations*, 2018 Proc, Int'l Conf. on Mgmt. Data 115.
- 235 See McSherry, *supra* note 229, at 91; Gaboardi et al., *supra* note 78, at 6.
- 236 See Kobbi Nissim, Sofya Raskhodnikova & Adam Smith, *Smooth Sensitivity and Sampling in Private Data Analysis*, 39 Proc. ACM Symp. on Theory Computing 75 (2007).
- 237 See Mohan et al., *supra* note 78, at 354; Roy et al., *supra* note 230.
- 238 See Gaboardi et al., *supra* note 78, at 21.
- 239 See *id.* at 2, 6.
- 240 See *id.* at 6.
- 241 *Harvard University Privacy Tools Project*, Harv. U., <https://privacytools.seas.harvard.edu/> [<https://perma.cc/ABN6-WVE3>] (last visited Oct. 1, 2018).
- 242 See Gaboardi et al., *supra* note 78, at 2.
- 243 See *id.*
- 244 See *id.*
- 245 See *id.* at 15, 19.
- 246 See *id.* at 2, 7.
- 247 See *id.* at 7.
- 248 See Xiaoqian Jiang et al., *A Community Assessment of Privacy Preserving Techniques for Human Genomes*, 14 BMC Med. Informatics & Decision Making 1, 1-2 (2014).
- 249 See Eland, *supra* note 226.
- 250 See Machanavajjhala et al., *supra* note 124, at 277.
- 251 See Darakhshan J. Mir et al., *DP-WHERE: Differentially Private Modeling of Human Mobility*, 2013 IEEE Int'l Conf. on Big Data 580, 580-82.
- 252 See Erlingsson et al., *RAPPOR*, *supra* note 216, at 1054.
- 253 See Greenberg, *supra* note 217.
- 254 See Gaboardi et al., *supra* note 78, at 6.
- 255 *Id.*
- 256 See Michael Hay et al., *Principled Evaluation of Differentially Private Algorithms Using DPBench*, 2016 Proc. ACM SIGMOD Int'l Conf. on Mgmt. Data 139, 139, <http://dl.acm.org/citation.cfm?id=2882931> [<https://perma.cc/6BQD-PQCT>].
- 257 *Id.*; see also DPComp, <https://www.dpcomp.org> [<https://perma.cc/72CL-86ZN>] (last visited Sept. 25, 2018).

- 258 See Ganta, Kasiviswanathan & Smith, *supra* note 129, at 265.
- 259 Here, the term “privacy attacks” refers to attempts to learn private information specific to individuals from a data release.
- 260 See *supra* note 84 and accompanying text. The term “magnitude” refers to the magnitude of the random noise distribution as measured in parameters like the standard deviation or variance. This is not necessarily referring to the magnitude of the actual random noise sampled from the noise distribution. Generally, greater uncertainty requires a larger noise magnitude.
- 261 More accurately, the noise  $Y$  is sampled from the Laplace distribution with a mean of 0 and standard deviation of  $\sqrt{2/\epsilon}$ . The exact shape of the noise distribution is important for proving that outputting  $m + Y$  preserves differential privacy, but can be ignored for the current discussion.
- 262 Note that this means that, when the sample size is small, the accuracy can be significantly reduced. For instance, if the sample size is similar in magnitude to  $1/\epsilon$ , the amount of noise that is added can even be larger than the sample size. Differential privacy works best when the sample size is large, specifically when it is significantly larger than  $1/\epsilon$ .
- 263 The standard deviation of the difference  $m - p \cdot n$  is  $\sqrt{p \cdot (1-p) / n}$  for small values of  $p$ . See Blitzstein & Hwang, *supra* note 102, at 158-60. Thus, the expected value of the deviation  $|m - p \cdot n|$  is approximately  $\sqrt{p \cdot (1-p) / n}$ . See J. Martin Bland & Douglas G. Altman, *Measuring Agreement in Method Comparison Studies*, 8 Stat. Methods Med. Res. 135, 147 (1999).
- 264 The expectation of  $m'$  is exactly  $m$  because the Laplace distribution has zero mean. The standard deviation of the difference  $m' - m$  is exactly the standard deviation of  $Y$ , which was chosen to be  $1/\epsilon$ .
- 265 Events are said to be statistically independent when the probability of occurrence of each event does not depend on whether the other event occurs. See Blitzstein & Hwang, *supra* note 102, at 56.
- 266 See Dwork & Roth, *supra* note 25, at 6, 22.
- 267 See *id.* at 20; Dwork et al., *supra* note 62, at 29; Vadhan, *supra* note 46, at 361.
- 268 See Dwork & Roth, *supra* note 25, at 192. When  $k$  is approximately  $1/\epsilon$ , the group privacy guarantee corresponds to  $k \cdot \epsilon \approx 1$ .
- 269 Guarantees that correspond to higher values than  $k \cdot \epsilon \approx 1$  (say,  $k \cdot \epsilon > 10$ ) provide only weak privacy guarantees.
- 270 See generally Dwork et al., *supra* note 62.



Try out [PMC Labs](#) and tell us what you think. [Learn More.](#)

**Gates Open Research**  
Immediate & Transparent Publishing



Version 2. [Gates Open Res.](#) 2019; 3: 1722.

PMCID: PMC7216402

Published online 2020 Apr 6. doi: [10.12688/gatesopenres.13089.2](https://doi.org/10.12688/gatesopenres.13089.2)

[Other versions](#)

PMID: [32478311](https://pubmed.ncbi.nlm.nih.gov/32478311/)

## Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff

[Samantha Petti](#), Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing<sup>1</sup> and [Abraham Flaxman](#), Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing<sup>a,2</sup>

<sup>1</sup>School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 30332, USA

<sup>2</sup>Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, 98121, USA

<sup>a</sup>Email: [abie@uw.edu](mailto:abie@uw.edu)

**Competing interests:** ADF has consulted recently for Kaiser Permanente; Sanofi; Merck for Mothers; Agathos, Ltd; and NORC. SP has no competing interests to disclose.

Accepted 2020 Mar 26.

[Copyright](#) : © 2020 Petti S and Flaxman A

This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Version Changes

---

#### Revised. Amendments from Version 1

This revised version improves the exposition in many places, thanks to helpful feedback from two peer reviewers, and adds detailed supplementary material on how empirical privacy loss (EPL) functions in the case of three simplified examples, including (1) when EPL is very close to epsilon; (2) when EPL is substantially less than epsilon, due to slack in the inequality in the Sequential Composition Theorem; and (3) when EPL is substantially more than epsilon, due to invariants.

#### Peer Review Summary

---

Review date	Reviewer name(s)	Version reviewed	Review status
<a href="#">2020 May 11</a>	David Van Riper		Approved
<a href="#">2020 May 6</a>	Ferdinando Fioretto		Approved
<a href="#">2020 Mar 3</a>	Ferdinando Fioretto		Approved with Reservations
<a href="#">2019 Dec 20</a>	David Van Riper		Approved with Reservations

## Abstract

---

**Background:** The 2020 US Census will use a novel approach to disclosure avoidance to protect respondents' data, called TopDown. This TopDown algorithm was applied to the 2018 end-to-end (E2E) test of the decennial census. The computer code used for this test as well as accompanying exposition has recently been released publicly by the Census Bureau.

**Methods:** We used the available code and data to better understand the error introduced by the E2E disclosure avoidance system when Census Bureau applied it to 1940 census data and we developed an empirical measure of privacy loss to compare the error and privacy of the new approach to that of a (non-differentially private) simple-random-sampling approach to protecting privacy.

**Results:** We found that the empirical privacy loss of TopDown is substantially smaller than the theoretical guarantee for all privacy loss budgets we examined. When run on the 1940 census data, TopDown with a privacy budget of 1.0 was similar in error and privacy loss to that of a simple random sample of 50% of the US population. When run with a privacy budget of 4.0, it was similar in error and privacy loss of a 90% sample.

**Conclusions:** This work fits into the beginning of a discussion on how to best balance privacy and accuracy in decennial census data collection, and there is a need for continued discussion.

**Keywords:** Decennial census, differential privacy, TopDown algorithm, empirical privacy loss

## Acronyms

---

DP - differentially private

E2E - end-to-end

TC - total count

SC - stratified count

MAE - median absolute error

EPL - empirical privacy loss

## Introduction

---

In the United States, the Decennial Census is an important part of democratic governance. Every ten years, the US Census Bureau is constitutionally required to count the “whole number of persons in each State,” and in 2020 this effort is likely to cost over 15 billion dollars <sup>1, 2</sup>. The results will be used for apportioning representation in the US House of Representatives and dividing federal tax dollars between states, as well as for a multitude of other governmental activities at the national, state, and local levels. Data from the decennial census will also be used extensively by sociologists, economists, demographers, and other researchers, and it will also inform strategic decisions in the private and non-profit sectors, and facilitate the accurate weighting of subsequent population surveys for the next decade <sup>3</sup>.

The confidentiality of information in the decennial census is also required by law, and the 2020 US Census will use a novel approach to “disclosure avoidance” to protect respondents’ data<sup>4</sup>. This approach builds on Differential Privacy, a mathematical definition of privacy that has been developed over the last decade and a half in the theoretical computer science and cryptography communities<sup>5</sup>. Although the new approach allows a more precise accounting of the variation introduced by the process, it also risks reducing the utility of census data—it may produce counts that are substantially less accurate than the previous disclosure avoidance system, which was based on redacting the values of table cells below a certain size (cell suppression) and a technique called swapping, where pairs of households with similar structures but different locations had their location information exchanged in a way that required that the details of the swapping procedure be kept secret<sup>6</sup>.

To date, there is a lack of empirical examination of the new disclosure avoidance system, but the approach was applied to the 2018 end-to-end (E2E) test of the decennial census, and computer code used for this test as well as accompanying exposition has recently been released publicly by the Census Bureau<sup>4,7</sup>.

We used the recently released code, preprints, and data files to understand and quantify the error introduced by the E2E disclosure avoidance system when the Census Bureau applied it to 1940 census data (for which the individual-level data has previously been released<sup>8</sup>) for a range of privacy loss budgets. We also developed an empirical measure of privacy loss and used it to compare the error and privacy of the new approach to that of a (non-differentially private) simple-random-sampling approach to protecting privacy.

## Methods

---

### Differential privacy definition and history

A randomized algorithm for analyzing a database is differentially private (DP) if withholding or changing one person’s data does not substantially change the algorithm’s output. If the results of the computation are roughly the same whether or not my data are included in the database, then the computation must be protecting my privacy. DP algorithms come with a parameter  $\epsilon$ , which quantifies how much privacy loss is allowed, meaning how much can one person’s data to affect the analysis.

To be precise, a randomized algorithm is  $\epsilon$ -DP if, for each possible output  $\mathcal{P}$ , for any pair of datasets  $D$  and  $D'$  that are the same everywhere except for on one person’s data,

$$\Pr [\mathcal{A}(D) = \mathcal{P}] \leq \exp(\epsilon) \Pr [\mathcal{A}(D') = \mathcal{P}].$$

Differential privacy is a characteristic of an algorithm; it is not a specific algorithm. Algorithms often achieve differential privacy by adding random variation<sup>5</sup>.

The new disclosure avoidance system for the 2020 US Census is designed to be DP and to maintain the accuracy of census counts. To complicate things beyond the typical challenge faced in DP algorithm design, there are certain counts in the census that will be published precisely as enumerated, without any variation added. These invariants have not been selected for the 2020 decennial census yet, but in the 2018 end-to-end (E2E) test, the total count for each state and the number of households in each enumeration district were invariants. There are also inequalities that will be enforced. The E2E test required the total count of people in an enumeration district to be greater or equal to the number of occupied households in that district<sup>9</sup>.

## TopDown algorithm

At a high level, the census approach to this challenge repeats two steps for multiple levels of a geographic hierarchy (from the top down, hence their name “TopDown”). The first step (Imprecise Histogram) adds variation from a carefully chosen distribution to the stratified counts of individuals. This produces a set of counts with illogical inconsistencies, which we refer to as an “imprecise histogram”. For example, counts in the imprecise histogram might be negative, might violate invariants or other inequalities, or might be inconsistent with the counts that are one level up in the geographic hierarchy. The second step (Optimize) finds optimized counts for each most-detailed cell in the histogram, using constrained convex optimization to make them as close as possible to the counts in the imprecise histogram, subject to the constraints that the optimized counts be non-negative, consistent with each other and the higher levels of the hierarchy, and satisfy the invariants and inequalities. These two steps are performed for each geographic level, from the coarsest to the finest. Each level is assigned a privacy budget  $\epsilon_i$  (which governs how much variation to add in the Imprecise Histogram step), and the entire algorithm achieves  $\epsilon$ -DP for  $\epsilon = \sum_i \epsilon_i$ . The 2020 US Census data may have six geographic levels, nested hierarchically: national, state, county, census tracts, block groups, and blocks; but in the 1940 E2E test four levels (national, state, county, and enumeration district) were included.

**Step one: Imprecise Histogram.** In the E2E algorithm applied to the 1940s microdata, TopDown added random variation in a flexible way that allowed the user to choose what statistics are the most important to keep accurate. The variation was added to the detailed histogram counts for the level and also to a preselected set of aggregate statistics. The detailed histogram counts stratified the population of each geographic by age (two values: under-18-year-olds and 18-plus), race (six values), ethnicity (two values: Hispanic and non-Hispanic), and household/group-quarters type (6 values). The aggregate statistics are sets of histogram count sums specified by some characteristics. For example, the “race/ethnicity/age” aggregate statistic contains 24 counts: people of each of the six racial categories who are also Hispanic ethnicity under age 18, of Hispanic ethnicity age 18 and over, of non-Hispanic ethnicity under age 18, and of non-Hispanic ethnicity age 18 and over.

The aggregate statistics (internally called “DP queries” in the TopDown algorithm) afford a way to choose specific statistics that are more important to keep accurate, and the E2E test included two such aggregates: a household/group-quarters query, which increases the accuracy of the count of each household type at each level of the hierarchy, and a race/ethnicity/age query, which increases the accuracy of the stratified counts of people by race, ethnicity, and voting age across all household/group-quarters types (again for each level of the spatial hierarchy). It also included “detailed queries” corresponding to boxes in the histogram. The detailed queries were afforded 10% of the privacy budget at each level, while the DP queries split the remaining 90% of the privacy budget, with 22.5% spent on the household/group-quarters queries and 67.5% spend on the race/ethnicity/age queries.

The epsilon budget of the level governed how much total random variation to add. A further parameterization of the epsilon budget determined how the variance was allocated between the histogram counts and each type of aggregate statistic. We write  $\epsilon_i = h + s_1 + s_2 + \dots + s_k$ , where  $\epsilon_i$  was the budget for the geographic level,  $h$  was the budget for the detailed queries, and  $s_1, \dots, s_k$  were the budgets for each of the  $k$  types of aggregate statistics. Then variance was added independently to each count according to the follow distribution:

$$\text{imprecise detailed histogram count} = \text{precise detailed histogram count} + G(h/2)$$

$$\text{imprecise aggregate stat } j = \text{precise aggregate stat } j + G(s_j/2)$$

where  $G(z)$  denotes the two-tailed geometric distribution,

$$\Pr [G(z) = k] = \frac{(1 - \exp(-z)) \exp(-zk)}{1 + \exp(-z)}.$$

The imprecise counts and imprecise aggregate statistics are unbiased estimates with variance  $(1 - \exp(-z))^2 / (2\exp(-z))$ , where  $z$  is the parameter for the geometric random variable added. A higher privacy budget means the variance added is more concentrated around zero, and therefore the corresponding statistic is more accurate. Therefore, adjusting the privacy budgets of the various aggregate statistics gives control over which statistics are the most private/least accurate (low fraction of the budget) and the most accurate/least private (high fraction of the budget).

The variation added to each histogram count comes from the same distribution, and is independent of all other added variation; the variance does not scale with the magnitude of count, e.g. adding 23 people to the count of age 18 and older non-Hispanic Whites is just as likely as adding 23 people to the count of age under 18 Hispanic Native Americans, even though the population of the latter is smaller.

**Step two: Optimize.** In this step, the synthetic data is created from the imprecise detailed histogram counts and aggregate statistics by optimizing a quadratic objective function subject to a system of linear equations and inequalities. The algorithm creates a variable for each detailed histogram count and each aggregate statistic. It adds equations and inequalities to encode the requirements that (i) each count and aggregate statistic is non-negative, (ii) the invariants and inequalities are satisfied, (iii) the aggregate statistics are the sum of the corresponding detailed histogram counts, and (iv) the statistics are consistent with the higher level synthetic data counts (i.e. the total number of people aged 18 and over summed across the counties in a state is equal to the number of people aged 18 and over in that state as reported by synthetic data set constructed in the previous phase). The optimization step finds a solution that satisfies these equations and minimizes the weighted sum of the squared differences between each variable/aggregate of variables and the corresponding imprecise detailed histogram count or imprecise aggregate statistic. This sum is weighted with the weight of each term taken to be proportional to the magnitude of the variation added in step one to create the imprecise count. The solution to this optimization is not necessarily integral, however, and TopDown uses a second optimization step to round fractional counts to integers.

We note that the approach that Census Bureau has taken with the TopDown where imprecise histogram data is optimized based on internal consistency has been developed in a line of research over the last decade to that has focused on obtaining count data that is DP *and* accurate [10-13](#).

### Empirical Privacy Loss for quantifying impact of optimize steps

As described above, the privacy loss of a DP algorithm is quantified by a unitless number,  $\epsilon$ , that bounds the maximum of the log of the relative change in the probability of an output when one person's data is changed. This bound is typically proven by logical deduction, and for complex DP algorithms, the proof often relies on the Sequential Composition Theorem [5](#), which states that information derived by combining the output of an  $\epsilon_1$ -DP algorithm and an  $\epsilon_2$ -DP algorithm is at most  $(\epsilon_1 + \epsilon_2)$ -DP. This theorem is an inequality, however, and the inequality might have room for improvement.



It is possible to empirically quantify privacy loss, which has the potential to show that the inequality of the sequential composition theorem is not tight. The brute force approach quantify privacy loss empirically is to search over databases  $D$  and  $D'$  that differ on one row to find the event  $E$  with the largest ratio of probabilities; this is too computationally intensive to be feasible for all but the simplest DP algorithms.

For algorithms that produce DP counts of multiple subpopulations, such as TopDown, it is possible to use the distribution of the residual difference between the precise count and the DP count to derive a proxy of the distribution produced by the brute force approach<sup>14</sup>. The special structure of count queries affords a way to avoid re-running the algorithm repeatedly, which is essential for TopDown, since it takes several hours to complete a single run of the algorithm. Assuming that the residual difference of the DP count minus the precise count is identically distributed for queries across similar areas (such as voting-age population across all enumeration districts), and then instead of focusing on only the histogram counts containing the individual who has changed, we used the residuals for all areal units to estimate the probability of the event we are after:

$$\Pr [\text{error}_j = k] \approx \left( \sum_{j'=1}^c 1 [\{\text{error}_{j'} = k\}] \right) / C =: \hat{p}_k,$$

where  $\text{error}_j$  is the residual difference of DP counts returned by TopDown minus the precise count for that same quantity in the 1940 census, and the  $\text{error}_{j'}$  are residuals for  $C$  other queries assumed to be exchangeable.

To measure the empirical privacy loss (EPL), we approximated the probability distribution of the residuals (DP count minus precise count at a selected level of the geographic hierarchy), which we denote  $p^{\text{KDE}}(x)$ , using Gaussian kernel density estimation (KDE) with a bandwidth of 0.1, and compare the log-ratio inspired by the definition of  $\epsilon$ -DP algorithms:

$$\text{EPL}(x) = \log \left( \frac{p^{\text{KDE}}(x)}{p^{\text{KDE}}(x+1)} \right);$$

$$\text{EPL} = \max_{x \in (-\infty, \infty)} \{\text{abs}(\text{EPL}(x))\}$$

See Supplementary Methods Appendix for additional detail on the design and validation of the EPL metric<sup>15</sup>.

### TopDown options still to be selected

---

There are seven key choices in implementing TopDown, that balance accuracy and privacy. We list them here, and state how they were set in the 2018 end-to-end test when run on the 1940s Census data:

1. Overall privacy. A range of  $\epsilon$  values, with {0.25, 0.50, 0.75, 1.0, 2.0, 4.0, 8.0} used in the E2E test run on the 1940 Census Data.

2. How to split this budget between national, state, county, tract, block group, and block. In the test run,  $\epsilon$  was split evenly between national, state, county, and enumeration district.
3. What aggregate statistics (also known as “DP Queries”) to include. In the test, two DP Queries were included: (i) counts stratified by age-group/race/ethnicity (and therefore aggregated over household/group-quarters type); and (ii) the household/group-quarters counts, which tally the total number of people living in each type of housing (in a household, in institutional facilities of certain types, in non-institutional facilities of certain types).
4. At each level, how to split level-budget between detailed queries and DP queries. The test run used 10% for detailed queries, 22.5% for household/group-quarters; and 67.5% for age-group-/race-/ethnicity-stratified counts.
5. What invariants to include. The test run held the total population count at the national and state level invariant.
6. What constraints to include. The test run constrained the total count of people to be greater or equal to total count of occupied households at each geographic level.
7. What to publish. The test run published a synthetic person file and synthetic household file for a range of  $\epsilon$  values, for four different seeds to the pseudorandom number generator.

### Our evaluation approach

1. We calculated residuals (DP count minus precise count) and summarized their distribution by its median absolute error (MAE) for total count (TC) and age/race/ethnicity stratified count (SC) at the state, county, and enumeration-district level. We also summarized the size of these counts from the precise-count versions to understand relative error as well as the absolute error introduced by TopDown.
2. We calculated a measure of empirical privacy loss (EPL), inspired by the definition of differential privacy. To measure EPL, we approximated the probability distribution of the residuals (DP count minus precise count at a selected level of the geographic hierarchy), which we denote  $p^{\text{KDE}}(x)$ , using Gaussian kernel density estimation with a bandwidth of 0.1, and compare the log-ratio inspired by the definition of  $\epsilon$ -DP algorithms:

$$\text{EPL}(x) = \log \left( \frac{p^{\text{KDE}}(x)}{p^{\text{KDE}}(x+1)} \right);$$

$$\text{EPL} = \max_{x \in (-\infty, \infty)} \{\text{abs}(\text{EPL}(x))\}$$

See Supplementary Methods Appendix for additional detail on the design and validation of the EPL metric [15](#). We hypothesized that the EPL of TopDown will be substantially smaller than the theoretical guarantee of  $\epsilon$ , which was proven using the Sequential Composition Theorem, which provides an inequality that is usually not a tight bound [14](#). However, it is possible that it will be much larger than  $\epsilon$ , due to the difficult-to-predict impact of including certain invariants.

3. We searched for bias in the residuals from (1), with our hypothesis that the DP counts are larger than precise counts in spatial areas with high homogeneity and DP counts are smaller than precise counts in areas with low homogeneity. We based this hypothesis on the expected impact of the non-negativity constraints included in the optimization steps of the TopDown algorithm. For each detailed query with a negative value for its noisy count, the optimization step will increase the value to make the results logical, and this reduction in variance must tradeoff some increase in bias. To quantify the scale of the bias introduced by optimization, for each geographic area, we constructed simple homogeneity index by counting the cells of the detailed histogram that contained a precise count of zero, and we examined the bias, defined as the mean of the DP count minus precise count, for these areas when stratified by homogeneity index.
4. We also compared the median absolute error and empirical privacy loss of TopDown to a simpler, but not-differentially-private approach to protecting privacy, Simple Random Sampling (i.e. sampling without replacement) for a range of sized samples. To do this, we generated samples without replacement of the 1940 Census Data for a range of sizes, and applied the same calculations from (1) and (2) to this alternatively perturbed data.

## Results

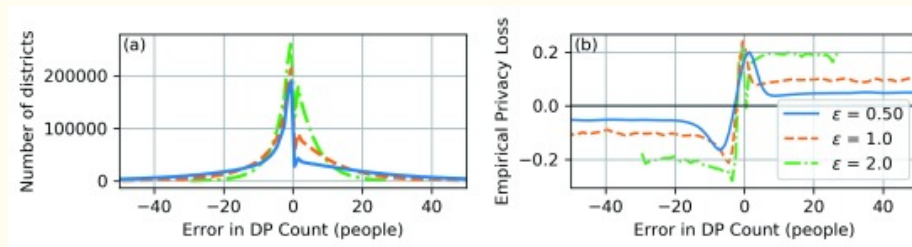
---

### Error and privacy of TopDown

Recall that geographic areas are nested: enumeration districts are contained within counties, which are contained within states. We found error in total count (TC) varied as a function of total privacy loss budget. Running TopDown with  $\epsilon = 0.5$  produced median absolute error in TC of 29 at the enumeration district level and 45 at the county level;  $\epsilon = 1.0$  produced median absolute error in TC of 15 at the enumeration district level and 24 at the county level; and  $\epsilon = 2.0$  produced median absolute error in TC of 8 at the enumeration district level and 13 at the county level (Full table in Extended Data [16](#)). At the state level, there was TC error of 0.0, as expected from the state TC invariant. The median and 95th percentile of TC from the precise-count data were 865 and 2342 for enumeration districts, 18,679 and 122,710 for counties, and 1,903,133 and 7,419,040 for states.

Error in stratified count (SC) varied similarly; when  $\epsilon = 0.5$ , the median absolute error in SC at the enumeration district level was 10 people, at the county level was 11 people, and at the state level was 13 people; for  $\epsilon = 1.0$ , the median absolute error in SC at the enumeration district level was 6 people, at the county level was 6 people, and at the state level was 7 people; and for  $\epsilon = 2.0$ , the median absolute error in SC at the enumeration district level was 4 people, at the county level was 4 people, and at the state level was 4 people. The median and 95th percentile of SC from the precise-count data were 88 and 967 for enumeration districts, 47 and 17,480 for counties, and 229 and 714,208 for states. (

[Figure 1](#))

**Figure 1.**

[Open in a separate window](#)

**Error distribution and empirical privacy loss for stratified counts at the enumeration district level.**

Panel (a) shows the distribution of residuals (DP - Precise) for stratified counts at the enumeration district level, stratified by age, race, and ethnicity; and panel (b) shows the empirical privacy loss function,  $EPL(x) = \log(\hat{p}^{KDE}(x)/\hat{p}^{KDE}(x+1))$ , where  $\hat{p}(x)$  is the probability density corresponding to the histogram in (a), after smoothing with a Gaussian kernel of bandwidth 0.1; the EPL value is the maximum of the absolute value of  $EPL(x)$  over all  $x$ .

We found that the empirical privacy loss was often substantially smaller than the privacy loss budget. For  $\epsilon = 0.5$ , the empirical privacy loss for TC at the enumeration district level was 0.033 and at the county level was 0.035 (at the state level empirical privacy loss is undefined, since the invariant makes all residuals zero); for  $\epsilon = 1.0$ , the empirical privacy loss for TC at the enumeration district level was 0.064 and at the county level was 0.048; and for  $\epsilon = 2.0$ , the empirical privacy loss for TC at the enumeration district level was 0.116 and at the county level was 0.094.

This relationship between privacy loss budget and empirical privacy loss was similar for stratified counts (SC) at the enumeration district and county level, but for privacy loss budgets of 1.0 and less, the empirical privacy at the enumeration district level was loss for SC was not as responsive to  $\epsilon$ . For  $\epsilon = 1.5$ , the empirical privacy loss for SC at the enumeration district level was 0.200, at the county level was 0.165, and at the state level was 0.104; for  $\epsilon = 1.0$ , the empirical privacy loss for SC at the enumeration district level was 0.241, at the county level was 0.164, and at the state level was 0.166; and for  $\epsilon = 2.0$ , the empirical privacy loss for SC at the enumeration district level was 0.280, at the county level was 0.253, and at the state level was 0.300. EPL values for all combinations of  $\epsilon$  and all geographic levels appear in the Extended Data.

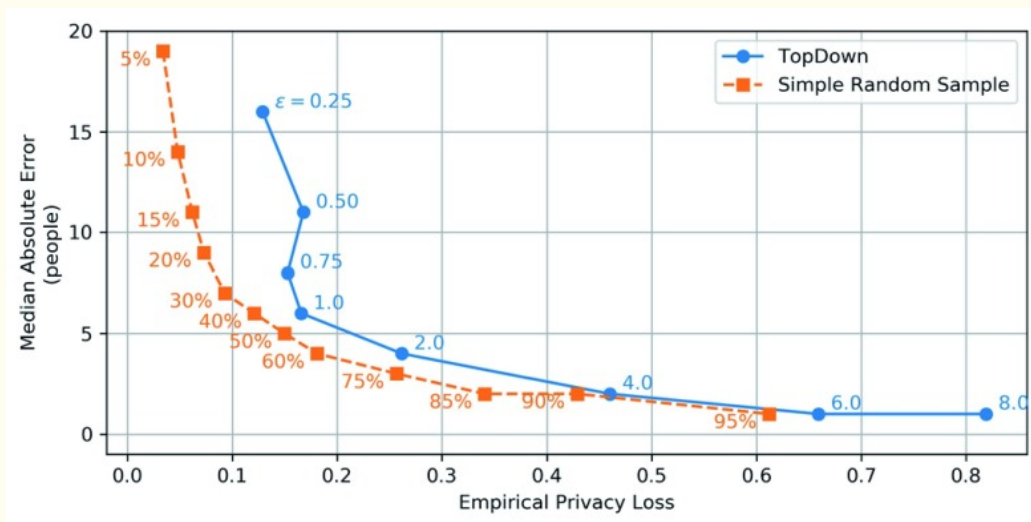
### Comparison with error and privacy of simple random sampling

We found that the MAE and EPL of Simple Random Sampling (i.e. sampling uniformly, without replacement) varied with larger sample size in a manner analogous to the total privacy budget in TopDown, for  $\epsilon \geq 1$ . For a 5% sample of the 1940 Census data, we found median absolute error in TC of 74 at the enumeration district level, 388 at the county level, and 3883 at the state level; a 50% sample produced median absolute error in TC of 17 at the enumeration district level, 90 at the county level, and 932 at the state level; and a 95% sample produced median absolute error in TC of 4 at the enumeration district level, 20 at the county level, and 130 at the state level.

Error in stratified count varied similarly; for a 5% sample, we found median absolute error in SC of 18 at the enumeration district level, 19 at the county level, and 41 at the state level; a 50% sample produced median absolute error in TC of 4 at the enumeration district level, 5 at the county level, and 9 at the state level.

We found empirical privacy loss increased as sample size increased. For a 5% sample, at the enumeration district level, we found EPL of 0.020 for TC and 0.098 for SC, and at the county level, we found 0.035 for TC and 0.034 for SC; a 50% sample produced EPL of 0.079 for TC and 0.318 for SC at the enumeration district level, and 0.082 for TC and 0.150 for SC at the county level; and a 95% sample produced EPL of 0.314 for TC and 1.333 for SC at the enumeration district level, and 0.429 for TC and 0.612 for SC at the county level ( [Figure 2](#), [Table 1](#)).

**Figure 2.**



[Open in a separate window](#)

Tradeoff curve of median absolute error and empirical privacy loss of stratified counts at the county level.

The curve with circular markers shows that in TopDown, the choice of  $\epsilon$  controls the tradeoff between MAE and EPL, although for  $\epsilon < 1$  there is not much difference in EPL. The curve with square markers shows the MAE and EPL of Simple Random Sampling for a range of sample sizes, for comparison. For example, TopDown with  $\epsilon = 1.0$ , provides privacy loss and estimation error similar to a sample of 50% of the 1940 census data, while  $\epsilon = 2.0$  is comparable to a 75% sample (for counts stratified by age, race, and ethnicity at the county level; different aggregate statistics produce different comparisons).

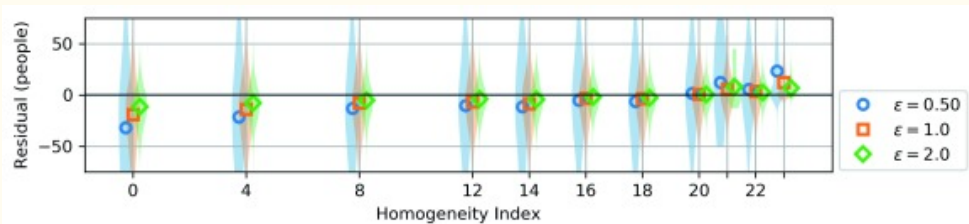
**Table 1.**

**Values of privacy loss, and corresponding proportions of Simple Random Sample (SRS) with most similar median-absolute-error/empirical-privacy-loss profile.**

Privacy Budget ( $\epsilon$ )	Closest SRS sample proportion (%)
1.0	50%
2.0	75%
4.0	90%
6.0	95%

### Bias in the variation introduced by TopDown

The bias introduced by TopDown varied with homogeneity index, as hypothesized. Enumeration districts with homogeneity index 0 (0 empty cells in the detailed histogram) had TC systematically lower than the precise count, while enumeration districts homogeneity index 22 (the maximum number of empty cells observed in the detailed histogram) had TC systematically higher than the precise count. The size of this bias decreased as a function of  $\epsilon$ . Homogeneity index 0 had bias of -31.7 people for  $\epsilon = 0.5$ , -18.9 people for  $\epsilon = 1.0$ , and -11.6 people for  $\epsilon = 2.0$ ; while homogeneity index 22 had bias of 5.4 people for  $\epsilon = 0.5$ , 3.6 people for  $\epsilon = 1.0$ , and 2.3 people for  $\epsilon = 2.0$ . ([Figure 3](#))

**Figure 3.**

[Open in a separate window](#)

#### Relationship between homogeneity index and residual for three values of epsilon.

The homogeneity index, defined as the number of cells with precise count of zero in the detailed histogram, is positively associated with the bias (markers show the mean difference between the DP count estimated by TopDown and the precise count, and shaded area shows the distribution of individual differences). This plot shows the association for enumeration districts, and a similar relationship holds at the county level. As  $\epsilon$  increases, the scale of the bias decreases. (Enumeration districts attained only a subset of the homogeneity index values between 0 and 23, which is why there are different width gaps between markers. We pooled the residuals for the four runs of TopDown with different random seed.)

Counties displayed the same general pattern, but there are fewer counties and they typically have less empty strata, so it was not as pronounced. The size of this bias again decreased as a function of  $\epsilon$ . Homogeneity index 0 had bias of -59.2 people for  $\epsilon = 0.5$ , -33.9 people for  $\epsilon = 1.0$ , and -18.8 people for  $\epsilon = 2.0$ ; while homogeneity index 22 had bias of 21.7 people for  $\epsilon = 0.5$ , 14.5 people for  $\epsilon = 1.0$ , and 11.1 people for  $\epsilon = 2.0$ .

## Discussion

---

We anticipate some readers of this will be social researchers who rely on Census Bureau data for quantitative work, and who have concerns that the Census Bureau is going to reduce the accuracy of this data. Such a reader may be open to the possibility that privacy is a valid reason for reducing accuracy, yet still be concerned about how this will affect their next decade of research. Our results visually summarized in [Figure 2](#) can help to understand the potential change in accuracy: if  $\epsilon = 1.0$ , for county-level stratified counts, TopDown will be like the uncertainty introduced by working with a 50% sample of the full dataset; if  $\epsilon = 2.0$ , it will be like working with a 75% sample; and if  $\epsilon = 6.0$ , it will have accuracy matching a 95% sample, which is pretty close to having the full data without protecting privacy. Such a reader may still want to see an analysis like this run on the 2010 decennial census data, but we hope this will help them rest a little easier about the quality of the data they are relying on for their work.

We also expect that some readers will be more drawn to the lower end of the epsilon curve. Just how private is TopDown with  $\epsilon = 0.25$ , especially when total count at the state-level is invariant? Our results show that all  $\epsilon$  less than 1.0 have empirical privacy loss around 0.15, independent of  $\epsilon$ . You can add more and more variation, but, perhaps due to the invariants, that variation does not translate into more and more privacy.

Comparing error in total count or stratified count across levels of the geographic hierarchy reveals a powerful feature of the TopDown algorithm: the error is of similar magnitude even though the counts are substantially different in size. This is because the variation added at each level has been specified to have the same portion of the total privacy budget. It remains to be investigated how alternative allocations of privacy budget across levels will change the error and empirical privacy loss.

For  $\epsilon \geq 1.0$ , TopDown introduced near minimal variation and attained empirical privacy loss almost 10 times less than  $\epsilon$ . We also found that this created a quantifiable amount of bias. The bias increased the reported counts in homogeneous districts while decreasing the counts in racially and ethnically mixed districts. The TopDown algorithm may therefore drive some small amount of redistribution of resources from diverse urban communities to segregated rural communities.

Accurate counts in small communities are important for emergency preparedness and other routine planning tasks performed by state and local government demographers, and this work may help to understand how such work will be affected by the shift to a DP disclosure avoidance system.

This work has not investigated more detailed research uses of decennial census data in social research tasks, such as segregation research, and how this may be affected by TopDown.

Another important use of decennial census data is in constructing control populations and survey weights for survey sampling of the US population for health, political, and public opinion polling. Our work provides some evidence on how TopDown may affect this application, but further work is warranted.

This work fits into the beginning of a discussion on how to best balance privacy and accuracy in decennial census data collection, and there is a need for continued discussion. This need must be balanced against a risky sort of observer bias—some researchers have hypothesized that calling attention to the privacy and confidentiality of census responses, even if done in a positive manner,

could reduce the willingness of respondents to answer census questions, and ongoing investigation with surveys and cognitive testing may provide some evidence on the magnitude of this effect as well as potential countermeasures <sup>17</sup>.

## Limitations

There are many differences between the 1940 census data and the 2020 data to be collected next year. In addition to the US population being three times larger now, the analysis will have six geographic levels instead of four, ten times more race groups and over 60 times more age groups. We expect that this will yield detailed queries with typical precise count sizes even smaller than the stratified counts for enumeration districts we have examined here. We suspect that impact of this will likely be to slightly decrease accuracy and increase privacy loss, but the accuracy of our hypothesis remains to be seen.

In addition to the changes in the data, additional changes are planned for TopDown, such as a switch from independent geometrically distributed variation to the High Dimensional Matrix Mechanism. We expect this to increase the accuracy a small amount without changing the empirical privacy loss.

In this work, we have focused on the median of the absolute error, but the spread of this distribution is important as well, and in future work, researchers may wish to investigate the tails of this distribution. We have also focused on the empirical privacy loss for specific queries at specific geographic aggregations, and our exploration was not comprehensive. Therefore, it is possible that some other test statistic would demonstrate a larger empirical privacy loss than we have found with our approach. Our approach also assumes that the residuals for different locations in a single run are an acceptable proxy for the residuals from the same location across multiple runs. Although these are certainly different, we suspect that the difference is sufficiently small as to not affect our estimates substantially.

## Conclusion

---

The TopDown algorithm will provide a provably  $\epsilon$ -DP disclosure avoidance system for the 2020 US Census, and it provides affordances to balances privacy and accuracy. This is an opportunity, but it is not without risks. Taking advantage of the opportunity and mitigating the risks will require that we understand what the approach is doing, and we hope that this analysis of the 2018 E2E test can help build such understanding.

## Data availability

---

### Source data

Individual-level data from the 1940 US Census is available from IPUMS <https://doi.org/10.18128/D010.V8.0.EXT1940USCB> <sup>8</sup>.

These data are under Copyright of Minnesota Population Center, University of Minnesota. Access to the documentation is freely available without restriction; however, users must register before extracting data from the website.

The output of the TopDown algorithm when run on the 1940 US Census data is available to download from the US Census Bureau: [https://www2.census.gov/census\\_1940/](https://www2.census.gov/census_1940/).

These data are under Copyright of the United States Census Bureau.

### Extended data



Zenodo: Extended data for Differential privacy in the 2020 US census, what will it do? Quantifying the accuracy/privacy tradeoff. <https://doi.org/10.5281/zenodo.3551215> <sup>16</sup>.

This project contains a full table of summary counts and errors for a range of levels of geographic hierarchy, stratification, and epsilon.

Zenodo: Supplementary Methods Appendix for Differential privacy in the 2020 US census, what will it do? Quantifying the accuracy/privacy tradeoff: Design and validation of Empirical Privacy Loss (EPL) metric. <https://doi.org/10.5281/zenodo.3727242> <sup>15</sup>.

This project contains additional details on the design and validation of the EPL metric used in this paper.

Extended data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

### Software availability

Scripts to produce all results and figures in this paper are available online: [https://github.com/aflaxman/dp\\_2020\\_census/](https://github.com/aflaxman/dp_2020_census/).

Archived scripts at time of publication: <https://doi.org/10.5281/zenodo.3551217> <sup>18</sup>.

License: [MIT License](#).

### Acknowledgements

---

Thanks to Neil Marquez (University of Washington) for suggesting comparing TopDown to simple random sampling. Thanks to danah boyd, Cynthia Dwork, Simson Garfinkel, Philip Leclerc, and Kunal Talwar for their helpful comments and discussion of this work.

### Notes

---

[version 2; peer review: 2 approved]

### Funding Statement

---

ADF is a recipient of funding from the Bill and Melinda Gates Foundation. The authors received no specific funding to support this work.

### References

---

1. Garfinkel S, Abowd JM, Martindale C: Understanding database reconstruction attacks on public data. *Communications of the ACM*. 2019;62(3):46–53. 10.1145/3287287 [[CrossRef](#)] [[Google Scholar](#)]
2. United States Government Accountability Office: Census Bureau improved the quality of its cost estimation but additional steps are needed to ensure reliability.U.S. G.A.O.2018. [Reference Source](#) [[Google Scholar](#)]
3. Ruggles S, Fitch C, Magnuson D, et al. : Differential privacy and Census data: Implications for social and economic research. *AEA papers and proceedings*. 2019;109:403–08. 10.1257/pandp.20191107 [[CrossRef](#)] [[Google Scholar](#)]
4. Abowd JM, Garfinkel SL: Disclosure avoidance and the 2018 Census test: Release of the source code. 2019. [Reference Source](#) [[Google Scholar](#)]

5. Dwork C, Roth A: The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* Now Publishers, Inc.2014;9(3–4):211–407. [Reference Source](#) [[Google Scholar](#)]
6. McKenna L: Disclosure avoidance techniques used for the 1970 through 2010 Decennial Censuses of Population and Housing. *Center for Economic Studies, U.S. Census Bureau*2018. [Reference Source](#) [[Google Scholar](#)]
7. boyd d: Differential privacy in the 2020 Decennial Census and the implications for available data products.CoRR abs/1907.03639.2019. [Reference Source](#) [[Google Scholar](#)]
8. Ruggles S, Flood S, Goeken R, et al. : IPUMS USA: Version 8.0 extract of 1940 Census for U.S. Census Bureau disclosure avoidance research [dataset]. 2018. 10.18128/D010.V8.0.EXT1940USCB [[CrossRef](#)] [[Google Scholar](#)]
9. Garfinkel S, others: 2018 end-to-end test disclosure avoidance system design specification.U.S. Census Bureau.2019. [Reference Source](#) [[Google Scholar](#)]
10. Hay M, Rastogi V, Miklau G, et al. : Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*. *VLDB Endowment*2010;1021–32. 10.14778/1920841.1920970 [[CrossRef](#)] [[Google Scholar](#)]
11. Li C, Miklau G, Hay M, et al. : The matrix mechanism: Optimizing linear counting queries under differential privacy. *The VLDB journal*. Springer. 2015;24:757–81. 10.1007/s00778-015-0398-x [[CrossRef](#)] [[Google Scholar](#)]
12. Kuo YH, Chiu CC, Kifer D, et al. : Differentially private hierarchical count-of-counts histograms. *Proceedings of the VLDB Endowment*. *VLDB Endowment*2018;1509–21. 10.14778/3236187.3236202 [[CrossRef](#)] [[Google Scholar](#)]
13. Fioretto F, Van Hentenryck P: Differential privacy of hierarchical census data: An optimization approach. *International conference on principles and practice of constraint programming* Springer.2019;639–55. 10.14778/3236187.3236202 [[CrossRef](#)] [[Google Scholar](#)]
14. Flaxman AD: Empirical quantification of privacy loss with examples relevant to the 2020 US Census. 2019. [Reference Source](#) [[Google Scholar](#)]
15. Flaxman AD, Petti S: Supplementary Methods Appendix to Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff: Design and validation of Empirical Privacy Loss (EPL) metric (Version v1.0).2020. 10.5281/zenodo.3727242 [[CrossRef](#)]
16. Petti S, Flaxman AD: Extended data for Differential privacy in the 2020 US census, what will it do? Quantifying the accuracy/privacy tradeoff [Data set]. *Zenodo*. 2020. 10.5281/zenodo.3718648 [[CrossRef](#)]
17. Childs JH, Abowd J: Update on confidentiality and disclosure avoidance.U.S. Census Bureau.2019. [Reference Source](#) [[Google Scholar](#)]
18. Petti S, Flaxman A: aflaxman/dp\_2020\_census: Replication archive code when paper was resubmitted (Version v1.0.1). *Zenodo*. 2020. 10.5281/zenodo.3718649 [[CrossRef](#)] [[Google Scholar](#)]

## Reviewer response for version 2

[David Van Riper](#), Referee<sup>1</sup>

<sup>1</sup>IPUMS, University of Minnesota, Minneapolis, MN, USA

**Competing interests:** No competing interests were disclosed.

Review date: 2020 May 11. Status: Approved. doi: [10.21956/gatesopenres.14301.r28749](https://doi.org/10.21956/gatesopenres.14301.r28749)

[Copyright](#) : © 2020 Van Riper D

This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

---

I am pleased with the authors' responses to my initial review. Adding extra data in the supplementary materials provides a fuller picture of the analyses they executed, and the clarifications added to the text enhance understanding. I also greatly appreciate the new Supplementary Methods Appendix that provides a more detailed discussion of the EPL measure.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Reviewer Expertise:

geography, demography, census data, differential privacy

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

## Reviewer response for version 2

[Ferdinando Fioretto](#), Referee<sup>1</sup>

<sup>1</sup>Syracuse University, Syracuse, NY, USA

**Competing interests:** No competing interests were disclosed.

Review date: 2020 May 6. Status: Approved. doi: [10.21956/gatesopenres.14301.r28748](https://doi.org/10.21956/gatesopenres.14301.r28748)

[Copyright](#) : © 2020 Fioretto F

This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

I am happy with the authors' response to my comments and with their new revised paper.

While I would have liked to see a more formal analysis and description of the method evaluated, I also understand the authors' desire to keep the article accessible to a wider audience.

To conclude, I believe that this article could be accepted without further revision.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Is the study design appropriate and is the work technically sound?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Reviewer Expertise:

Artificial Intelligence, Differential Privacy, Optimization

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

## Reviewer response for version 1

[Ferdinando Fioretto](#), Referee<sup>1</sup>

<sup>1</sup>Syracuse University, Syracuse, NY, USA

**Competing interests:** No competing interests were disclosed.

Review date: 2020 Mar 3. Status: Approved with Reservations. doi: [10.21956/gatesopenres.14238.r28430](https://doi.org/10.21956/gatesopenres.14238.r28430)

[Copyright](#) : © 2020 Fioretto F

This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

---

### Overview

The paper examines the behavior of TopDown, a privacy-preserving algorithm proposed to release differentially private US Census data. The authors examine the privacy, accuracy, and bias trade-off induced by the application of TopDown on the 1940 US Census dataset. The analysis was detailed for various privacy loss levels (i.e., epsilon values) and compared against a simple random sampling approach.

The authors provide a brief overview of Differential Privacy and the TopDown algorithm. Next, they introduce the empirical privacy loss as an empirical quantification of the loss of privacy induced by the application of a differentially private mechanism, and, finally, they provide an extensive evaluation on an application of TopDown on the 1940 US Census data release.

An interesting aspect of this work is the introduction of a novel evaluation metric, called "empirical privacy loss" or EPL. The authors argue that the use of the post-processing strategy adopted by TopDown, that projects the differentially private solution into a feasible space, may reduce the theoretical privacy loss and the experimental evaluation seem to support such claim. In particular, the authors found that the EPL for a given class of counts (total count and stratified count) is smaller than the theoretical privacy loss guaranteed by the algorithm. I have several comments about this metric, reported in the detailed comments section.

I found this work original, in that it provides an extensive evaluation of the privacy, accuracy, and bias trade-off of the Top-Down algorithm. However, I also found the absence of a related work section unusual and would like to point out that there are other works that use optimization techniques to publish accurate count statistics, e.g.:

- Michael Hay, Vibhor Rastogi, Gerome Miklau, Dan Suciu: Boosting the Accuracy of Differentially Private Histograms Through Consistency. PVLDB 3(1): 1021-1032 (2010) <sup>1</sup>.
- Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, Vibhor Rastogi: The matrix mechanism: optimizing linear counting queries under differential privacy. VLDB J. 24(6): 757-781 (2015) <sup>2</sup>.

and work that pose particular emphasis on Census data:

- Yu-Hsuan Kuo, Cho-Chun Chiu, Daniel Kifer, Michael Hay, Ashwin Machanavajjhala: Differentially Private Hierarchical Count-of-Counts Histograms. PVLDB 11(11): 1509-1521 (2018) <sup>3</sup>.
- Ferdinando Fioretto, Pascal Van Hentenryck: Differential Privacy of Hierarchical Census Data: An Optimization Approach. CP 2019: 639-655 <sup>4</sup>.

It may be useful to discuss some of these proposals.

The paper is well organized and described with a good amount of detail. However, I would have liked to see a more formal description of the TopDown algorithm and of the empirical privacy loss concept. In particular, I believe that describing TopDown using an optimization model would greatly simplify readability and avoid some doubts, such as those I list in my detailed comments. I would also suggest the authors introduce an illustration of the hierarchy utilized by the Census, together with the amount of privacy budget used at each level. This could, for instance, be visualized as a tree, where the root node describes the total counts at the national level, its children describe counts at the state level, and so on. I believe that such an illustration will ease visualizing the process performed by TopDown during Step 2, in order to satisfy the consistency of the problem constraints.

It would also be useful to have a table summarizing the problem constraints. For example, the authors describe equalities constraints, such as those that constrain the aggregate statistics and counts as well as those that force the invariants, and inequality constraints, such as non-negativity and properties over the group sizes.

#### **Detailed Comments:**

#### **Section: TopDown algorithm**

- The authors provide a helpful overview of the TopDown algorithm, which operates in two steps: Noise addition and Optimization. I believe that the description can be further improved--I found the text to be quite verbose--and would encourage the authors to supply the following information:
  - A table that summarizes the attributes of the histograms to be produced (e.g., counts of each geographic by age, race, ethnicity, household/group quarters) and the aggregate statistics.
  - An illustration highlighting the dependence between counts, and, thus, the constraints arising from these dependencies.

I believe the above can be a helpful aid in the description of the algorithm.

- The authors call "aggregate statistics" as "DP queries". I am not sure why this terminology was selected. At the best of my knowledge, a DP query is simply a function over a dataset that happens to satisfy DP. I would suggest using a different terminology for identifying private aggregates.
- At the end of the third paragraph of **Step One: Imprecise Histogram**: I would have preferred to see a more formal description for the computation of the histogram count and aggregate statistics. For instance, in the current version, it is not clear what is the dimensionality of each query.
- In **Step two: Optimize**: the authors describe how TopDown optimizes the noisy estimates to satisfy the problem constraints. I would strongly suggest using a mathematical model to describe the problem (minimizer and constraints). In the current stage, a reader unfamiliar with the topic may find some sentences confusing. For example, the sentence "finds a solution that [...] has the property that the value of each variable is as close as possible to the corresponding imprecise detailed histogram count or imprecise aggregate statistics" may denote that the objective is to minimize some  $L_p$  distance between the optimized counts and noisy ones; but for which  $p$ ? I think that adding a formal model would improve the paper clarity.

### Section: Empirical Privacy Loss

- I found the introduction of the empirical privacy loss concept quite interesting. However, I also have a few reservations. First, I think that the formula in this section could be described in more detail. I may have missed something, but I could not find what  $C$  correspond to. Also, this formula seems to be hard to compute and I wish the authors have spent a few words on they address such a challenge.
- The notation  $\hat{p}_k$  used in the formula  $\Pr[\text{error} \dots]$  seems to have the same semantic of notation  $\hat{p}(x)$ , introduced in point (2) of Section "**Our evaluation approach**". Is this correct, i.e., is it that  $\hat{p}_k = \hat{p}(k)$ ? If this is the case, then one of the two notations need to be changed for consistency.
- In section **TopDown Options still to be selected**:
  - On point (1): I suggest spacing the epsilon values listed;
  - On point (4): I wonder if the authors have some intuitions on why the test run used more budget for aggregated statistics than for aggregated queries. I believe it would be very insightful to discuss the implications of such budget partitioning.
- In section **Our evaluation approach**:

- Point (2): I would have liked if the authors could have further elaborated on how the empirical privacy loss is computed. Is it the maximum among all  $x$  of  $ELP(x)$ ?
- The authors specify that the EPL is computed for the total count and they report a substantially lower loss than the theoretical privacy budget adopted. Since the privacy budget was partitioned among several levels and queries, I wonder if the authors have taken such partitioning into account when computing the final EPL score. I believe this aspect should be discussed in the text.
- Have the authors validated the fidelity of the EPL score on a simple differential privacy application? For instance, I would have liked to see a brief discussion on if this metric is in agreement with the theoretical errors provided by the Laplace mechanism on counting queries (without post-processing).

## Results

### Error and privacy of TopDown

- The authors explain in detail the results attained in their analysis. I found the reporting of the results at the end of each subsection to be a bit distracting. I suggest the authors introduce one or multiple tables that tabulate the results and only summarize them in the text.
- Additionally, the plots in Figure 1 and the errors describes in the text are for different privacy budget: The figure illustrates the errors for  $\epsilon = 0.5, 1.0, \text{ and } 2.0$ , while the text describes the errors for  $\epsilon = 0.25, 1.0, \text{ and } 4.0$ . I suggest the authors reporting the results for all the  $\epsilon$  tested into a table, or to make the description in the text and the figure consistent for the privacy budgets adopted.
- The empirical privacy loss computed was reported for the total count at the enumeration district level and country-level and compared against the privacy budget adopted by the TopDown algorithm. As stated in my comment above, I wonder if this comparison is fair. TopDown seems to partition the privacy budget for different queries, thus leaving the total count queries with substantially less budget than the original total one. I encourage the author to expand on this aspect of the evaluation.

### Comparison with error and privacy of simple random sampling

- As for the previous section, I recommend the authors to use a table to tabulate the numerical results described in the last paragraph. In my opinion, it will substantially increase readability.

### Bias in the variation introduced by TopDown

- As for the previous section, I suggest the authors tabulate the results of the homogeneity index and bias.
- Are the errors by homogeneity index an average over the sample runs?

Is the work clearly and accurately presented and does it cite the current literature?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Is the study design appropriate and is the work technically sound?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Reviewer Expertise:

Artificial Intelligence, Differential Privacy, Optimization

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

## References

---

1. : Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*.2010;3(1-2) : 10.14778/1920841.1920970 1021-1032 10.14778/1920841.1920970 [[CrossRef](#)] [[Google Scholar](#)]
2. : The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB Journal*.2015;24(6) : 10.1007/s00778-015-0398-x 757-781 10.1007/s00778-015-0398-x [[CrossRef](#)] [[Google Scholar](#)]
3. : Differentially private hierarchical count-of-counts histograms. *Proceedings of the VLDB Endowment*.2018;11(11) : 10.14778/3236187.3236202 1509-1521 10.14778/3236187.3236202 [[CrossRef](#)] [[Google Scholar](#)]
4. : Differential Privacy of Hierarchical Census Data: An Optimization Approach.2019;11802: 10.1007/978-3-030-30048-7\_37 639-655 10.1007/978-3-030-30048-7\_37 [[CrossRef](#)] [[Google Scholar](#)]

## Reviewer response for version 1

[David Van Riper](#), Referee<sup>1</sup>

<sup>1</sup>PUMS, University of Minnesota, Minneapolis, MN, USA

**Competing interests:** No competing interests were disclosed.

Review date: 2019 Dec 20. Status: Approved with Reservations. doi: [10.21956/gatesopenres.14238.r28334](https://doi.org/10.21956/gatesopenres.14238.r28334)

[Copyright](#) : © 2019 Van Riper D

This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

---

## Overview



Using differentially private 1940 census data produced by the US Census Bureau's TopDown algorithm, Petti and Flaxman assess the privacy/accuracy trade-off along multiple dimensions for this algorithm for multiple values of epsilon. The authors analyzed the median absolute error, empirical privacy loss, and bias for the differentially private data. They also compared the median absolute error and empirical privacy loss for differentially private data with data generated through simple random sampling. This is one of the first, if not the first, article assessing the accuracy of decennial census data published through a differentially private algorithm.

Petti and Flaxman provide a good overview of differential privacy and the Census Bureau's TopDown algorithm - a differentially private algorithm for producing decennial census data. They then compare the differentially private 1940 data with the original complete-count 1940 data to assess the accuracy introduced by the TopDown algorithm. They find that error increased as the total privacy loss budget decreased. They also find that empirical privacy loss was smaller than total privacy loss budget. They measure bias introduced by the algorithm and find that bias increases as homogeneity decreases and that bias increases as total privacy loss budget decreases. They conclude that privacy loss does not vary much for  $\epsilon < 1.0$ , and that the accuracy achieved when using a 50% simple random sample is equivalent to an epsilon of 1.0.

I am intrigued by the empirical privacy loss measure introduced by Petti and Flaxman. Its formula and interpretation mirrors the formula for epsilon-differential privacy. However, I would like to see a more thorough discussion of empirical privacy loss summary statistic reported in the results section of the paper. The authors compare an empirical privacy loss summary statistic with total privacy loss budget on pages 6 and 7 of the paper, but they never explain how the summary statistic was computed. Having that explanation would help me better understand the comparison they make throughout the paper.

The authors compare the empirical privacy loss for a given geographic unit-type of count (total count, stratified count) combination with the overall privacy loss budget. They empirical privacy loss for a given combination is less than the overall privacy loss budget. I wonder if this is the correct comparison to make. The privacy loss budget controls the overall amount of privacy leaked by the publication of all statistics. It is the sum, via sequential composition, of the epsilon fractions assigned to each geographic level-statistic combination. Thus, by definition, the empirical privacy loss associated with a particular geographic level-statistic (e.g., total population count) must be less than the privacy loss budget. I would like to see a fuller discussion of this comparison in the paper. See detailed comment #14 for more details.

I would also additional supplemental datasets (or tables in the paper) with the empirical privacy loss summary statistics for all values of epsilon. The authors report a few values in the text and figures, but having a complete set would allow for a more comprehensive understanding of the relationship between empirical privacy loss and epsilon.

Finally, I strongly recommend that the authors use the same examples in their text as they use in the figures. The text uses epsilons of 0.25, 1.0 and 4.0 and the figures use epsilons of 0.50, 1.0, and 2.0. Making the epsilons consistent between the text and figures will help the reader better understand the analysis.

### *Detailed comments by section*

#### **Methods - TopDown algorithm**

- The authors' high level overview (first paragraph in subsection entitled "TopDown algorithm") describe the noise injection (Imprecise Histogram) and optimization steps in the TopDown algorithm. They state that the "second step (Optimize) adjusts the histogram to be close as possible to the imprecise counts". I am uncertain about what histogram the authors refer to in this

sentence. Is the histogram based on the original data, or is this the noise-injected detailed histogram? My understanding of the algorithm is that it generates histograms (one for each combination of geographic level and query) from the original data and then injects noise into histograms using the appropriate two-sided geometric distribution. It then passes these noise-injected histograms to the optimization function.

I would like the authors to be more precise in their description of the histogram and the "imprecise counts" in this section.

- The authors state that the 2020 US Census will have six geographic levels nested hierarchically (last sentence of TopDown algorithm paragraph). The Census Bureau allocated privacy loss budget to seven nested geographies (nation, state, county, tract group, census tract, block group, block) for the 2010 demonstration product. The Bureau has not committed to this allocation for 2020 and could still change the allocation strategy. I recommend clarifying that statement to pertain solely to the 2010 demonstration data product.
- In the final clause of the last sentence of the TopDown algorithm paragraph, the authors state that "in the 1940 E2E test, only national, state, county, and district levels were included." I recommend adding the word "enumeration" before district in that clause.

#### **Methods - Step one: Imprecise Histogram**

- At the end of first paragraph in this section, the authors describe the "ethnicity-age" aggregate statistic set. The implication of this sentence is that the "ethnicity-age" aggregate statistics set was one pre-selected by Census for noise injection. Census did not choose this aggregate statistic set. The aggregate statistic sets chosen by census were Voting age by Hispanic origin by Race (a 2 x 2 x 6 cell query) and Household/Group quarter (a 6 cell query). I recommend modifying this sentence to describe one of the two pre-selected aggregate statistic sets.
- At the end of the second paragraph, the authors write that "22.5% spent on the group-quarters queries". I recommend changing the fragment to be "22.5% spent on the household/group-quarters queries". The word "household" is important when discussing this DP query. People can either live in household or group quarters, and by definition, households are not group quarters.

#### **Methods - TopDown options still to be selected**

- For option 3, I recommend modifying the "(and therefore aggregated over "group quarters types)" to be "(there therefore aggregated over "household/group quarters types)". A household is not a type of group quarter.
- Also in option 3, I recommend modifying the "(ii) the group-quarters counts" to be "(ii) the household/group quarters counts".
- In option 5, add the word "population" between "total" and "count" in the second sentence. Otherwise, readers will not necessarily know which total count to which the authors are referring.

#### **Results - Error and privacy of TopDown**

- At the end of first paragraph of this subsection, the authors list the median and 95th percentile of TC for EDs, counties, and states. I think it is important to clarify that these counts are based on the original 1940 census data and not on any of the differentially private 1940 datasets. Since this

sentence comes at the end of a paragraph describing median absolute error, readers may assume the medians and 95th percentiles are from a DP dataset. Consider moving that sentence up the start of the paragraph.

- At the end of second paragraph of this subsection, the authors list the median and 95th percentile of SC for EDs, counties, and states. I think it is important to clarify that these counts are based on the original 1940 census data and not on any of the differentially private 1940 datasets. Since this sentence comes at the end of a paragraph describing median absolute error, readers may assume the medians and 95th percentiles are from a DP dataset. Consider moving that sentence up the start of the paragraph.
- The final two paragraphs of this subsection describe the empirical privacy loss for TC and SC for different geographic levels and different epsilons. They describe the EPL for epsilons of 0.25, 1.0, and 4.0 in the text. I would like to have a table, either in the paper or in the extended data product, that lists the EPLs for all values of epsilon and all geographic levels for TC and SC. I wonder how linear the relationship between EPL and epsilon is.
- The authors list a number of EPL values in the final two paragraphs and in the right-hand panel of Figure 1, but I do not know what the EPL value represents. Is it the absolute value of the maximum observed EPL, or is it the range from the maximum to minimum observed EPL value? I would appreciate a more complete discussion of how the authors calculated the value of EPL they plot in Figure 1 and list in the text. The formula on page 5 describes how to compute EPL for a single geographic unit and value of epsilon, but I don't see how that formula extends to the summary statistics reported on page 6.
- Figure 1 plots the error and EPL for epsilon equal to 0.5, 1.0, and 2.0, but the text in the final two paragraphs describes EPL for epsilons of 0.25, 1.0, and 4.0. I strongly recommend making the values in the text and the plot consistent with one another. That consistency will make it easier to interpret the plot in Figure 1.
- The authors compare the empirical privacy loss for a given geographic unit-type of count (total count, stratified count) combination with the overall privacy loss budget. They empirical privacy loss for a given combination is less than the overall privacy loss budget. I wonder if this is the correct comparison to make. The privacy loss budget controls the overall amount of privacy leaked by the publication of all statistics. It is the sum, via sequential composition, of the epsilon fractions assigned to each geographic level-statistic combination. Thus, by definition, the empirical privacy loss associated with a particular geographic level-statistic (e.g., total population count) must be less than the privacy loss budget.

For a given value of epsilon, we can compute the portion of that value that is assigned to each geographic level - query combination. For example, epsilon of 0.25 is divided up as follows:

Geographic levels = 0.25 to each level

Tables = 0.1 (detailed), 0.225 (household-group quarters), 0.675 (voting age - Hispanic - race)

We can multiply the geographic level fraction by the table fractions by epsilon to yield:

Geog level - detailed query = 0.00625 epsilon

Geog level - household group quarters query = 0.0140625 epsilon

Geog level - voting age - Hispanic - race query = 0.0421875 epsilon

These epsilons still do not equate to an epsilon associated with a particular statistic, such as total population count. Given the optimization step and the state-level total population invariant, I'm not sure if we can compute an epsilon value for a particular statistic. But these epsilon values seems like a more appropriate comparison to the empirical privacy loss reported by the authors.

### **Results - Comparison with error and privacy of simple random sampling**

- I would like to have a table of MAE and EPL values for Simple Random Sampling. Consider adding those values to the Extended Data product currently available, or adding another Extended Data product with these values.
- Consider adding a plot of EPL by sample size to supplement or even replace the final paragraph of this subsection. There are a lot of numbers in the final paragraph, and I find it difficult to visualize the relationship between EPL and sampling fraction just by reading the numbers.
- The x-axis for Figure 2 depicts values of Empirical Privacy Loss, but neither the text nor the caption describe how the values were computed. This comment fits with comment 12 - what does the Empirical Privacy Loss summary statistic mean and how was it computed.

### **Results - Bias in the variation introduced by TopDown**

- Figure 3 plots the error and EPL for epsilon equal to 0.5, 1.0, and 2.0, but the text in the first paragraph describes EPL by homogeneity index for epsilons of 0.25, 1.0, and 4.0. I strongly recommend making the values in the text and the plot consistent with one another. That consistency will make it easier to interpret the plot in Figure 3.
- I recommend moving the (Figure 3) parenthetical to the end of the discussion on EPL by homogeneity for enumeration districts. Figure 3 only shows the results for enumeration districts, but the parenthetical comes after the discussion for counties.
- In the paragraph and Figure 3, the authors list a summary statistic for bias by homogeneity index and epsilon. Is the summary statistic the mean or the median?
- Figure 3 displays the violin plot/mean bias for 11 of 23 homogeneity index values. I recommend modifying the figure caption to indicate that the authors are only displaying some of the homogeneity index values on the plot.
- I also recommend modifying the x-axis label to indicate that the homogeneity index values are for enumeration districts. That would help readers immediately understand what geographic units are being plotted.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Reviewer Expertise:

geography, demography, census data, differential privacy

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

Articles from Gates Open Research are provided here courtesy of **Gates Foundation - Open Access**

---

# A Firm Foundation for Private Data Analysis

Cynthia Dwork  
Microsoft Research  
dwork@microsoft.com

## 1. PRIVATE DATA ANALYSIS

In the information realm, loss of privacy is usually associated with failure to control access to information, to control the flow of information, or to control the purposes for which information is employed. *Differential privacy* arose in a context in which ensuring privacy is a challenge even if all these control problems are solved: privacy-preserving statistical analysis of data.

The problem of *statistical disclosure control* – revealing accurate statistics about a set of respondents while preserving the privacy of individuals – has a venerable history, with an extensive literature spanning statistics, theoretical computer science, security, databases, and cryptography (see, for example, the excellent survey [1], the discussion of related work in [2] and the *Journal of Official Statistics* 9(2), dedicated to confidentiality and disclosure control). This long history is a testament the importance of the problem. Statistical databases can be of enormous social value; they are used for apportioning resources, evaluating medical therapies, understanding the spread of disease, improving economic utility, and informing us about ourselves as a species.

The data may be obtained in diverse ways. Some data, such as census, tax, and other sorts of official data, are compelled; others are collected opportunistically, for example, from traffic on the internet, transactions on Amazon, and search engine query logs; other data are provided altruistically, by respondents who hope that sharing their information will help others to avoid a specific misfortune, or more generally, to increase the public good. Altruistic data donors are typically promised their individual data will be kept confidential – in short, they are promised “privacy.” Similarly, medical data and legally compelled data, such as census data, tax return data, have legal privacy mandates. In our view, ethics demand that opportunistically obtained data should be treated no differently, especially when there is no reasonable alternative to engaging in the actions that generate the data in question.

The problems remain: even if data encryption, key management, access control, and the motives of the data curator

are all unimpeachable, what does it mean to preserve privacy, and how can it be accomplished?

### 1.1 “How” is Hard

Let us consider a few common suggestions and some of the difficulties they can encounter.

*Large Query Sets.* One frequent suggestion is to disallow queries about a specific individual or small set of individuals. A well-known differencing argument demonstrates the inadequacy of the suggestion. Suppose it is known that Mr. X is in a certain medical database. Taken together, the answers to the two large queries “How many people in the database have the sickle cell trait?” and “How many people, not named X, in the database have the sickle cell trait?” yield the sickle cell status of Mr. X. The example also shows that encrypting the data, another frequent suggestion (oddly), would be of no help at all. The privacy compromise arises from correct operation of the database.

In *query auditing* each query to the database is evaluated in the context of the query history to determine if a response would be disclosive; if so, then the query is refused. For example, query auditing might be used to interdict the pair of queries about sickle cell trait just described. This approach is problematic for several reasons, among them that query monitoring is computationally infeasible [15] and that the refusal to respond to a query may itself be disclosive [14].

We think of a database as a collection of *rows*, with each row containing the data of a different respondent. In *subsampling* a subset of the rows is chosen at random and released. Statistics can then be computed on the subsample and, if the subsample is sufficiently large, these may be representative of the dataset as a whole. If the size of the subsample is very small compared to the size of the dataset, this approach has the property that every respondent is unlikely to appear in the subsample. However, this is clearly insufficient: Suppose appearing in a subsample has terrible consequences. Then every time subsampling occurs *some* individual suffers horribly.

In *input perturbation*, either the data or the queries are modified before a response is generated. This broad category encompasses a generalization of subsampling, in which the curator first chooses, based on a secret, random, function of the query, a subsample from the database, and then returns the result obtained by applying the query to the subsample [4]. A nice feature of this approach is that repeating the same query yields the same answer, while semantically equivalent but syntactically different queries are made on essentially unrelated subsamples. However, an outlier may only be protected by the unlikelihood of being in the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

subsample.

In what is traditionally called *randomized response*, the data themselves are randomized once and for all and statistics are computed from the noisy responses, taking into account in the distribution on the perturbation [22]. The term “randomized response” comes from the practice of having the respondents to a survey flip a coin and, based on the outcome, answering an invasive yes/no question or answering a more emotionally neutral one. In the computer science literature the choice governed by the coin flip is usually between honestly reporting one’s value and responding randomly, typically by flipping a second coin and reporting the outcome. Randomized response was devised for the setting in which the individuals do not trust the curator, so we can think of the randomized responses as simply being published. Privacy comes from the uncertainty of how to interpret a reported value. The approach becomes untenable for complex data.

*Adding random noise to the output* has promise, and we will return to it later; here we point out that if done naïvely this approach will fail. To see this, suppose the noise has mean zero and that fresh randomness is used in generating every response. In this case, if the same query is asked repeatedly, then the responses can be averaged, and the true answer will eventually emerge. This is disastrous: an adversarial analyst could exploit this to carry out the difference attack described above. The approach cannot be “fixed” by recording each query and providing the same response each time a query is re-issued. There are several reasons for this. For example, syntactically different queries may be semantically equivalent, and, if the query language is sufficiently rich, then the equivalence problem itself is undecidable, so the curator cannot even test for this.

Problems with noise addition arise even when successive queries are completely unrelated to previous queries [5]. Let us assume for simplicity that the database consists of a single – but very sensitive – bit per person, so we can think of the database as an  $n$ -bit Boolean vector  $d = (d_1, \dots, d_n)$ . This is an abstraction of a setting in which the database rows are quite complex, for example, they may be medical records, but the attacker is interested in one specific field, such as HIV status. The abstracted attack consists of issuing a string of queries, each described by a subset  $S$  of the database rows. The query is asking how many 1’s are in the selected rows. Representing the query as the  $n$ -bit characteristic vector of the set  $S$ , with 1’s in all the positions corresponding to rows in  $S$  and 0’s everywhere else, the true answer to the query is the inner product  $A(S) = \sum_{i=1}^n d_i S_i$ . Suppose the privacy mechanism responds with  $A(S) + \text{random noise}$ . How much noise is needed in order to preserve privacy?

Since we have not yet defined privacy, let us consider the easier problem of avoiding blatant “non-privacy,” defined as follows: the system is blatantly non-private if an adversary can construct a candidate database that agrees with the real database  $D$  in, say, 99% of the entries. An easy consequence of the following theorem is that a privacy mechanism adding noise with magnitude always bounded by, say,  $n/401$  is blatantly non-private against an adversary that can ask all  $2^n$  possible queries [5]. There is nothing special about 401; any number exceeding 400 would work.

**THEOREM 1. [5]** *Let  $\mathcal{M}$  be a mechanism that adds noise bounded by  $E$ . Then there exists an adversary that can re-*

*construct the database to within  $4E$  positions.*

Blatant non-privacy with  $E = n/401$  follows immediately from the theorem, as the reconstruction will be accurate in all but at most  $4E = n \cdot \frac{4}{401} < n/100$  positions.

**PROOF.** Let  $d$  be the true database. The adversary can attack in two phases:

1. **Estimate the number of 1’s in all possible sets:** Query  $\mathcal{M}$  on all subsets  $S \subseteq [n]$ .
2. **Rule out “distant” databases:** For every candidate database  $c \in \{0, 1\}^n$ , If, for any  $S \subseteq [n]$ ,  $|\sum_{i \in S} c_i - \mathcal{M}(S)| > E$ , then rule out  $c$ . If  $c$  is not ruled out, then output  $c$  and halt.

Since  $\mathcal{M}(S)$  never errs by more than  $E$ , the real database will not be ruled out, so this simple (but inefficient!) algorithm will output *some* database; let us call it  $c$ . We will argue that the number of positions in which  $c$  and  $d$  differ is at most  $4 \cdot E$ .

Let  $I_0$  be the indices in which  $d_i = 0$ , that is,  $I_0 = \{i \mid d_i = 0\}$ . Similarly, define  $I_1 = \{i \mid d_i = 1\}$ . Since  $c$  was not ruled out,  $|\mathcal{M}(I_0) - \sum_{i \in I_0} c_i| \leq E$ . However, by assumption  $|\mathcal{M}(I_0) - \sum_{i \in I_0} d_i| \leq E$ . It follows from the triangle inequality that  $c$  and  $d$  differ in at most  $2E$  positions in  $I_0$ ; the same argument shows that they differ in at most  $2E$  positions in  $I_1$ . Thus,  $c$  and  $d$  agree on all but at most  $4E$  positions.  $\square$

What if we consider more realistic bounds on the number of queries? We think of  $\sqrt{n}$  as an interesting threshold on noise, for the following reason: if the database contains  $n$  people drawn uniformly at random from a population of size  $N \gg n$ , and the fraction of the population satisfying a given condition is  $p$ , then we expect the number of rows in the database satisfying  $p$  to be roughly  $np \pm \Theta(\sqrt{n})$ , by the properties of the hypergeometric distribution. That is, the sampling error is on the order of  $\sqrt{n}$ . We would like that the noise introduced for privacy is smaller than the sampling error, ideally  $o(\sqrt{n})$ . Unfortunately, noise of magnitude  $o(\sqrt{n})$  is blatantly non-private against a series of  $n \log^2 n$  *randomly generated* queries [5], no matter the distribution on the noise. Several strengthenings of this pioneering result are now known. For example, if the entries in  $S$  are chosen independently according to a standard normal distribution, then blatant non-privacy continues to hold even against an adversary asking only  $\Theta(n)$  questions, and even if more than a fifth of the responses have arbitrarily wild noise magnitudes, provided the other responses have noise magnitude  $o(\sqrt{n})$  [8].

These are not just interesting mathematical exercises. We have been focussing on *interactive* privacy mechanisms, distinguished by the involvement of the curator in answering each query. In the *noninteractive* setting the curator publishes some information of arbitrary form, and the data are not used further. Research statisticians like to “look at the data,” and we have frequently been asked for a method of generating a “noisy table” that will permit highly accurate answers to be derived for computations that are not specified at the outset. The noise bounds say this is impossible: no such table can safely provide very accurate answers to too many weighted subset sum questions; otherwise the table could be used in a simulation of the interactive mechanism,

and an attack could be mounted against the table. Thus, even if the analyst only requires the responses to a small number of unspecified queries, the fact that the table can be exploited to gain answers to other queries is problematic.

In the case of “internet scale” data sets, obtaining responses to, say,  $n \geq 10^8$  queries is infeasible. What happens if the curator permits only a sublinear number of questions? This inquiry led to the first algorithmic results in differential privacy, in which it was shown how to maintain privacy against a sublinear number of *counting* queries, that is, queries of the form “How many rows in the database satisfy property  $P$ ?” by adding noise of order  $o(\sqrt{n})$  – less than the sampling error – to each answer [11]. The cumbersome privacy guarantee, which focused on the question of what an adversary can learn about a row in the database, is now known to imply a natural and still very powerful relaxation of differential privacy.

## 1.2 “What” is Hard

Newspaper horror stories about “anonymized” and “de-identified” data typically refer to non-interactive approaches in which certain kinds of information in each data record have been suppressed or altered. A famous example is AOL’s release of a set of “anonymized” search query logs. People search for many “obviously” disclosive things, such as their full names (“vanity searches”), their own social security numbers (to see if their numbers are publicly available on the web, possibly with a goal of detecting assess the threat of identity theft), and even the combination of mother’s maiden name and social security number. AOL carefully redacted such obviously disclosive “personally identifiable information,” and each user id was replaced by a random string. However, search histories can be very idiosyncratic, and a New York Times reporter correctly traced such an “anonymized” search history to a specific resident of Georgia.

In a *linkage attack*, released data are linked to other databases or other sources of information. We use the term *auxiliary information* to capture information about the respondents *other* than that which is obtained through the (interactive or non-interactive) statistical database. Any priors, beliefs, or information from newspapers, labor statistics, and so on, all fall into this category.

In a notable demonstration of the power of auxiliary information, medical records of the governor of Massachusetts were identified by linking voter registration records to “anonymized” Massachusetts Group Insurance Commission (GIC) medical encounter data, which retained the birthdate, sex, and zip code of the patient [21].

Despite this exemplary work, it has taken several years to fully appreciate the importance of taking auxiliary information into account in privacy-preserving data release. Sources and uses of auxiliary information are endlessly varied. As a final example, it has been proposed to modify search query logs by mapping *all* terms, not just the user ids, to random strings. In *token-based hashing* each query is tokenized, and then an uninvertible hash function is applied to each token. The intuition is that the hashes completely obscure the terms in the query. However, using a statistical analysis of the hashed log and *any* (unhashed) query log, for example, the released AOL log discussed above, the anonymization can be severely compromised, showing that token-based hashing is unsuitable for anonymization [16].

As we will see next, there are deep reasons for the fact that auxiliary information plays such a prominent role in these examples.

## 2. DALENIUS’S DESIDERATUM

In 1977 the statistician Tore Dalenius articulated an “*ad omnia*” (as opposed to *ad hoc*) privacy goal for statistical databases: anything that can be learned about a respondent from the statistical database should be learnable without access to the database. Although informal, this feels like the “right” direction. The breadth of the goal captures all the common intuitions for privacy. In addition, the definition only holds the database accountable for whatever “extra” is learned about an individual, beyond that which can be learned from other sources. In particular, an extrovert who posts personal information on the web may destroy her own privacy, and the database should not be held accountable.

Formalized, Dalenius’ goal is strikingly similar to the gold standard for security of a cryptosystem against a passive eavesdropper, defined 5 years later. *Semantic security* captures the intuition that the encryption of a message reveals no information about the message. This is formalized by comparing the ability of a computationally efficient adversary, having access to both the ciphertext and any auxiliary information, to output (anything about) the plaintext, to the ability of a computationally efficient party having access *only* to the auxiliary information (and not the ciphertext), to achieve the same goal [12]. Abilities are measured by probabilities of success, where the probability space is over the random choices made in choosing the encryption keys, the ciphertexts, and by the adversaries. Clearly, if this difference is very, very tiny, then in a rigorous sense the ciphertext leaks (almost) no information about the plaintext.

The formal definition of semantic security is a pillar of modern cryptography. It is therefore natural to ask whether a similar property, such as Dalenius’ goal, can be achieved for statistical databases. But there is an essential difference in the two problems. Unlike the eavesdropper on a conversation, the statistical database attacker is also a user, that is, a legitimate consumer of the information provided by the statistical database (not to mention the fact that she may also be a respondent in the database).

Many papers in the literature attempt to formalize Dalenius’ goal (in some cases unknowingly) by requiring that the adversary’s prior and posterior views about an individual (*i.e.*, before and after having access to the statistical database) shouldn’t be “too different,” or that access to the statistical database shouldn’t change the adversary’s views about any individual “too much.” The difficulty with this approach is that if the statistical database teaches us anything at all, then it *should* change our beliefs about individuals. For example, suppose the adversary’s (incorrect) prior view is that everyone has 2 left feet. Access to the statistical database teaches that almost everyone has one left foot and one right foot. The adversary now has a very different view of whether or not any given respondent has two left feet. But has privacy been compromised?

The last hopes for Dalenius’ goal evaporate in light of the following parable, which again involves auxiliary information. Suppose we have a statistical database that teaches average heights of population subgroups, and suppose further that it is infeasible to learn this information (perhaps for financial reasons) in any other way (say, by conducting a



new study). Finally, suppose that one’s true height is considered sensitive. Given the auxiliary information “Turing is two inches taller than the average Lithuanian woman,” access to the statistical database teaches Turing’s height. In contrast, anyone without access to the database, knowing only the auxiliary information, learns much less about Turing’s height.

A rigorous impossibility result generalizes this argument, extending to essentially any notion of privacy compromise, *assuming the statistical database is useful*. The heart of the attack uses extracted randomness from the statistical database as a one-time pad for conveying the privacy compromise to the adversary/user [6].

Turing did not have to be a member of the database for the attack described above to be prosecuted against him. More generally, the things that statistical databases are designed to teach can, sometimes indirectly, cause damage to an individual, even if this individual is not in the database.

In practice, statistical databases are (typically) created to provide some anticipated social gain; they teach us something we could not (easily) learn without the database. Together with the attack against Turing described above, and the fact that he did not have to be a member of the database for the attack to work, this suggests a new privacy goal: minimize the increased risk to an individual incurred by joining (or leaving) the database. That is, we move from comparing an adversary’s prior and posterior views of an individual to comparing the risk to an individual when included in, versus when not included in, the database. This makes sense. A privacy guarantee that limits risk incurred by joining therefore encourages participation in the dataset, increasing social utility. This is the starting point on our path to *differential privacy*.

### 3. DIFFERENTIAL PRIVACY

Differential privacy will ensure that ability of an adversary to inflict harm (or good, for that matter) – of any sort, to any set of people – should be essentially the same, independent of whether any individual opts in to, or opts out of, the dataset. We will do this indirectly, simultaneously addressing all possible forms of harm and good, by focussing on the probability of any given output of a privacy mechanism and how this probability can change with the addition or deletion of any row. Thus, we will concentrate on pairs of databases ( $D, D'$ ) differing only in one row, meaning one is a subset of the other and the larger database contains just one additional row. Finally, to handle worst case pairs of databases, our probabilities will be over the random choices made by the privacy mechanism.

**DEFINITION 2.** *A randomized function  $\mathcal{K}$  gives  $\epsilon$ -differential privacy if for all data sets  $D$  and  $D'$  differing on at most one row, and all  $S \subseteq \text{Range}(\mathcal{K})$ ,*

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{K}(D') \in S], \quad (1)$$

where the probability space in each case is over the coin flips of  $\mathcal{K}$ .

The multiplicative nature of the guarantee implies that an output whose probability is zero on a given database must also have probability zero on any neighboring database, and hence, by repeated application of the definition, on any other database. Thus, Definition 1 trivially rules out the

subsample-and-release paradigm discussed above: for an individual  $x$  not in the dataset, the probability that  $x$ ’s data are sampled and released is obviously zero; the multiplicative nature of the guarantee ensures that the same is true for an individual whose data *are* in the dataset.

Any mechanism satisfying this definition addresses all concerns that any participant might have about the leakage of her personal information, regardless of any auxiliary information known to an adversary: even if the participant removed her data from the data set, no outputs (and thus consequences of outputs) would become significantly more or less likely. For example, if the database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of *any* individual’s data in the database will not significantly affect her chance of receiving coverage.

Definition 2 extends naturally to group privacy. repeated application of the definition bounds the ratios of probabilities of outputs when a collection  $C$  of participants opts in or opts out by a factor of  $e^{|C|\epsilon}$ . Of course, the point of the statistical database is to disclose aggregate information about large groups (while simultaneously protecting individuals), so we should expect privacy bounds to disintegrate with increasing group size.

The parameter  $\epsilon$  is public, and its selection is a social question. We tend to think of  $\epsilon$  as, say, 0.01, 0.1, or in some cases,  $\ln 2$  or  $\ln 3$ .

Sometimes, for example, in the census, an individual’s participation is known, so hiding presence or absence makes no sense; instead we wish to hide the values in an individual’s row. Thus, we can (and sometimes do) extend “differing in at most one row” to mean having symmetric difference at most 1 to capture both possibilities. However, we will continue to use Definition 2.

Returning to randomized response, we see that it yields  $\epsilon$ -differential privacy for a value of  $\epsilon$  that depends on the universe from which the rows are chosen and the probability with which a random, rather than non-random, value is contributed by the respondent. As an example, suppose each row consists of a single bit, and that the respondent’s instructions are to first flip an unbiased coin to determine whether he will answer randomly or truthfully. If heads (respond randomly), then the respondent is to flip a second unbiased coin and report the outcome; if tails, the respondent answers truthfully. Fix  $b \in \{0, 1\}$ . If the true value of the input is  $b$ , then  $b$  is output with probability 3/4. On the other hand, if the true value of the input is  $1 - b$ , then  $b$  is output with probability 1/4. The ratio is 3, yielding  $(\ln 3)$ -differential privacy.

Suppose  $n$  respondents each employ randomized response independently, but using coins of known, fixed, bias. Then, given the randomized data, by the properties of the binomial distribution the analyst can approximate the true answer to the question “How many respondents have value  $b$ ?” to within an expected error on the order of  $\Theta(\sqrt{n})$ . As we will see, it is possible to do much better – obtaining *constant* expected error, independent of  $n$ .

Generalizing in a different direction, suppose each row now has two bits, each one randomized independently, as described above. While each bit remains  $(\ln 3)$ -differentially private, their logical-AND enjoys less privacy. That is, consider a privacy mechanism in which each bit is protected by this exact method of randomized response, and consider the

query: “What is the logical-AND of the bits in the row of respondent  $i$  (after randomization)?” If we consider the two extremes, one in which respondent  $i$  has data 11 and the other in which respondent  $i$  has data 00, we see that in the first case the probability of output 1 is 9/16, while in the second case the probability is 1/16. Thus, this mechanism is at best  $(\ln 9)$ -differentially private, not  $\ln 3$ . Again, it is possible to do much better, even while releasing the entire 4-element histogram, also known as a *contingency table*, with only constant expected error in each cell.

#### 4. ACHIEVING DIFFERENTIAL PRIVACY

Achieving differential privacy revolves around hiding the presence or absence of a single individual. Consider the query “How many rows in the database satisfy property  $P$ ?” The presence or absence of a single row can affect the answer by at most 1. Thus, a differentially private mechanism for a query of this type can be designed by first computing the true answer and then adding random noise according to a distribution with the following property:

$$\forall z, z' \text{ s.t. } |z - z'| = 1 : \Pr[z] \leq e^\epsilon \Pr[z']. \quad (2)$$

To see why this is desirable, consider any feasible response  $r$ . For any  $m$ , if  $m$  is the true answer and the response is  $r$  then the random noise must have value  $r - m$ ; similarly, if  $m - 1$  is the true answer and the response is  $r$ , then the random noise must have value  $r - m + 1$ . In order for the response  $r$  to be generated in a differentially private fashion, it suffices for

$$e^{-\epsilon} \leq \frac{\Pr[\text{noise} = r - m]}{\Pr[\text{noise} = r - m + 1]} \leq e^\epsilon.$$

In general we are interested in vector-valued queries; for example, the data may be points in  $\mathbf{R}^d$  and we wish to carry out an analysis that clusters the points and reports the location of the largest cluster.

**DEFINITION 3.** [7] For  $f : \mathcal{D} \rightarrow \mathbf{R}^d$ , the  $L_1$  sensitivity of  $f$  is

$$\begin{aligned} \Delta f &= \max_{D, D'} \|f(D) - f(D')\|_1 \\ &= \max_{D, D'} \sum_{i=1}^d |f(D)_i - f(D')_i| \end{aligned} \quad (3)$$

for all  $D, D'$  differing in at most one row.

In particular, when  $d = 1$  the sensitivity of  $f$  is the maximum difference in the values that the function  $f$  may take on a pair of databases that differ in only one row. This is the difference our noise must be designed to hide. For now, let us focus on the case  $d = 1$ .

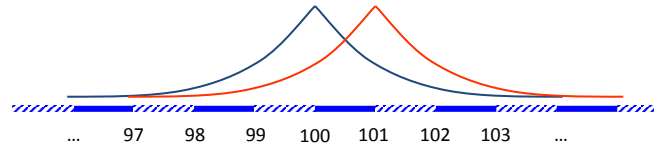
The Laplace distribution with parameter  $b$ , denoted  $\text{Lap}(b)$ , has density function  $P(z|b) = \frac{1}{2b} \exp(-|z|/b)$ ; its variance is  $2b^2$ . Taking  $b = 1/\epsilon$  we have that the density at  $z$  is proportional to  $e^{-\epsilon|z|}$ . This distribution has highest density at 0 (good for accuracy), and for any  $z, z'$  such that  $|z - z'| \leq 1$  the density at  $z$  is at most  $e^\epsilon$  times the density at  $z'$ , satisfying the condition in Equation 2. It is also symmetric about 0, and this is important. We cannot, for example, have a distribution that only yields non-negative noise. Otherwise the only databases on which a counting query could return a response of 0 would be databases in which no row satisfies the query. Letting  $D$  be such a database, and letting

$D' = D \cup \{r\}$  for some row  $r$  satisfying the query, the pair  $D, D'$  would violate  $\epsilon$ -differential privacy. Finally, the distribution gets flatter as  $\epsilon$  decreases. This is correct: smaller  $\epsilon$  means better privacy, so the noise density should be less “peaked” at 0 and change more gradually as the magnitude of the noise increases.

There is nothing special about the cases  $d = 1, \Delta f = 1$ :

**THEOREM 4.** [7] For  $f : \mathcal{D} \rightarrow \mathbf{R}^d$ , the mechanism  $\mathcal{K}$  that adds independently generated noise with distribution  $\text{Lap}(\Delta f/\epsilon)$  to each of the  $d$  output terms enjoys  $\epsilon$ -differential privacy.

Before proving the theorem, we illustrate the situation for the case of a counting query ( $\Delta f = 1$ ) when  $\epsilon = \ln 2$  and the true answer to the query is 100. The distribution on the outputs (in gray) is centered at 100. The distribution on outputs when the true answer is 101 is shown in orange.



**PROOF.** (Theorem 4.) The proof is a simple generalization of the reasoning we used to illustrate the case of a single counting query.

Consider any subset  $S \subseteq \text{Range}(\mathcal{K})$ , and let  $D, D'$  be any pair of databases differing in at most one row. When the database is  $D$ , the probability density at any  $r \in S$  is proportional to  $\exp(-\|f(D) - r\|_1(\epsilon/\Delta f))$ . Similarly, when the database is  $D'$ , the probability density at any  $r \in \text{Range}(\mathcal{K})$  is proportional to  $\exp(-\|f(D') - r\|_1(\epsilon/\Delta f))$ .

We have

$$\begin{aligned} \frac{e^{-\|f(D) - r\|_1(\epsilon/\Delta f)}}{e^{-\|f(D') - r\|_1(\epsilon/\Delta f)}} &= \frac{e^{\|f(D') - r\|_1(\epsilon/\Delta f)}}{e^{\|f(D) - r\|_1(\epsilon/\Delta f)}} \\ &= e^{(\|f(D') - r\|_1 - \|f(D) - r\|_1)(\epsilon/\Delta f)} \\ &\leq e^{(\|f(D') - f(D)\|_1)(\epsilon/\Delta f)} \end{aligned}$$

where the inequality follows from the triangle inequality. By definition of sensitivity,  $\|f(D') - f(D)\|_1 \leq \Delta f$ , and so the ratio is bounded by  $\exp(\epsilon)$ . Integrating over  $S$  yields  $\epsilon$ -differential privacy.  $\square$

Given any query sequence  $f_1, \dots, f_m$ ,  $\epsilon$ -differential privacy can be achieved by running  $\mathcal{K}$  with noise distribution  $\text{Lap}(\sum_{i=1}^m \Delta f_i/\epsilon)$  on each query, even if the queries are chosen adaptively, with each successive query depending on the answers to the previous queries. In other words, by allowing the quality of each answer to deteriorate in a controlled way with the sum of the sensitivities of the queries, we can maintain  $\epsilon$ -differential privacy.

With this in mind, let us return to some of the suggestions we considered earlier. Recall that using the specific randomized response strategy described above, for a single Boolean attribute, yielded error  $\Theta(\sqrt{n})$  on databases of size  $n$  and  $(\ln 3)$ -differential privacy. In contrast, using Theorem 4 with the same value of  $\epsilon$ , noting that  $\Delta f = 1$  yields a variance of  $2(1/\ln 3)^2$ , or an expected error of  $\sqrt{2}/\ln 3$ . More generally, to obtain  $\epsilon$ -differential privacy we get an expected error of  $\sqrt{2}/\epsilon$ . Thus, our expected error magnitude is constant, independent of  $n$ .

What about two queries? The sensitivity of a sequence of two counting queries is 2. Applying the theorem with

$\Delta f/\varepsilon = 2/\varepsilon$ , adding independently generated noise distributed as  $\text{Lap}(2/\varepsilon)$  to each true answer yields  $\varepsilon$ -differential privacy. The variance is  $2(2/\varepsilon)^2$ , or standard deviation  $2\sqrt{2}/\varepsilon$ . Thus, for any desired  $\varepsilon$  we can achieve  $\varepsilon$ -differential privacy by increasing the expected magnitude of the errors as a function of the total sensitivity of the two-query sequence. This holds equally for

- Two instances of the *same query*, addressing the repeated query problem;
- One count for each of two different bit positions, for example, when each row consists of two bits;
- A pair of queries of the form: “How many rows satisfy property P?” and “How many rows satisfy property Q?” (where possibly  $P = Q$ ); and
- An arbitrary pair of queries.

However, the theorem also shows we can sometimes do better. The logical-AND count we discussed above, even though it involves two different bits in each row, still only has sensitivity 1: the number of 2-bit rows whose entries are both 1 can change by at most one with the addition or deletion of a single row. Thus, this more complicated query can be answered in an  $\varepsilon$ -differentially private fashion using noise distributed as  $\text{Lap}(1/\varepsilon)$ ; we don’t need to use the distribution  $\text{Lap}(2/\varepsilon)$ .

### Histogram Queries.

The power of Theorem 4 really becomes clear when considering *histogram queries*, defined as follows. If we think of the rows of the database as elements in a universe  $X$ , then a histogram query is a partitioning of  $X$  into an arbitrary number of disjoint regions  $X_1, X_2, \dots, X_d$ . The implicit question posed by the query is: “For  $i = 1, 2, \dots, d$ , how many points in the database are contained in  $X_i$ ?” For example, the database may contain the annual income for each respondent, the query is a partitioning of incomes into ranges:  $\{[0, 50K), [50K, 100K), \dots, \geq 500K\}$ . In this case  $d = 11$ , and the question is asking, for each of the  $d$  ranges, how many respondents in the database have annual income in the given range. This looks like  $d$  separate counting queries, but the entire query actually has sensitivity  $\Delta f = 1$ . To see this, note that if we remove one row from the database, then only one cell in the histogram changes, and that cell only changes by 1; similarly for adding a single row. So Theorem 4 says that  $\varepsilon$ -differential privacy can be maintained by perturbing each cell with an independent random draw from  $\text{Lap}(1/\varepsilon)$ . Returning to our example of two-bit rows, we can pose the 4-ary histogram query requesting, for each pair of literals  $v_1v_2$ , the number of rows with value  $v_1v_2$ , adding noise of order  $1/\varepsilon$  to each of the four cells.

### When Noise Makes No Sense.

There are times when the addition of noise for achieving privacy makes no sense. For example, the function  $f$  might map databases to strings, strategies, or trees, or it might be choosing the “best” among some specific, not necessarily continuous, set of real-valued objects. The problem of optimizing the output of such a function while preserving  $\varepsilon$ -differential privacy requires additional technology.

Assume the curator holds a database  $D$  and the goal is to produce an object  $y$ . The *exponential mechanism* [18] works

as follows. We assume the existence of a *utility function*  $u(D, y)$  that measures the quality of an output  $y$ , given that the database is  $D$ . For example, the data may be a set of labeled points in  $\mathbf{R}^d$  and the output  $y$  might be a  $d$ -ary vector describing a  $(d-1)$ -dimensional hyperplane that attempts to classify the points, so that those labeled with +1 have non-negative inner product with  $y$  and those labeled with -1 have negative inner product. In this case the utility would be the number of points correctly classified, so that higher utility corresponds to a better classifier. The exponential mechanism,  $\mathcal{E}$ , outputs  $y$  with probability proportional to  $\exp(u(D, y)\varepsilon/\Delta u)$  and ensures  $\varepsilon$ -differential privacy. Here  $\Delta u$  is the sensitivity of the utility function bounding, for all adjacent databases  $(D, D')$  and potential outputs  $y$ , the difference  $|u(D, y) - u(D', y)|$ . In our example,  $\Delta u = 1$ . The mechanism assigns most mass to the best classifier, and the mass assigned to any other drops off exponentially in the decline in its utility for the current data set – hence the name “exponential mechanism.”

### When Sensitivity is Hard to Analyze.

The Laplace and exponential mechanisms provide a differentially private interface through which the analyst can access the data. Such an interface can be useful even when it is difficult to determine the sensitivity of the desired function or query sequence; it can also be used to run an iterative algorithm, composed of easily analyzed steps, for as many iterations as a given privacy budget permits. This is a powerful observation; for example, using only noisy sum queries, it is possible to carry out many standard datamining tasks, such as singular value decompositions, finding an ID3 decision tree, clustering, learning association rules, and learning anything learnable in the statistical queries learning model, frequently with good accuracy, in a privacy-preserving fashion [2]. This approach has been generalized to yield a publicly available codebase for writing programs that ensure differential privacy [17].

### k-Means Clustering.

As an example of “private programming” [2], consider  $k$ -means clustering, described first in its usual, non-private form. The input consists of points  $p_1, \dots, p_n$  in the  $d$ -dimensional unit cube  $[0, 1]^d$ . Initial candidate means  $\mu_1, \dots, \mu_k$  are chosen randomly from the cube and updated as follows:

1. Partition the samples  $\{p_i\}$  into  $k$  sets  $S_1, \dots, S_k$ , associating each  $p_i$  with the nearest  $\mu_j$ .
2. For  $1 \leq j \leq k$ , set  $\mu'_j = \sum_{i \in S_j} p_i / |S_j|$ , the mean of the samples associated with  $\mu_j$ .

This update rule is typically iterated until some convergence criterion has been reached, or a fixed number of iterations have been applied.

Although computing the nearest mean of any one sample (Step 1) would breach privacy, we observe that to compute an average among an unknown set of points it is enough to compute their sum and divide by their number. Thus, the computation only needs to expose the approximate cardinalities of the  $S_j$ , not the sets themselves. Happily, the  $k$  candidate means implicitly define a histogram query, since they partition the space  $[0, 1]^d$  according to their Voronoi cells, and so the vector  $(|S_1|, \dots, |S_k|)$  can be released with

very low noise in each coordinate. This gives us a differentially private approximation to the denominators in Step 2. As for the numerators, the sum of a subset of the  $p_i$  has sensitivity at most  $d$ , since the points come from the bounded region  $[0, 1]^d$ . Even better, the sensitivity of the  $d$ -ary function that returns, for each of the  $k$  Voronoi cells, the  $d$ -ary sum of the points in the cell is at most  $d$ : adding or deleting a single  $d$ -ary point can affect at most one sum, and that sum can change by at most 1 in each of the  $d$  dimensions. Thus, using a query sequence with total sensitivity at most  $d+1$ , the analyst can compute a new set of candidate means by dividing, for each  $\mu_j$ , the approximate sum of the points in  $S_j$  by the approximation to the cardinality  $|S_j|$ .

If we run the algorithm for a fixed number  $N$  of iterations we can use the noise distribution  $\text{Lap}((d+1)N/\varepsilon)$  to obtain  $\varepsilon$ -differentially privacy. If we don't know the number of iterations in advance we can increase the noise parameter as the computation proceeds. There are many ways to do this. For example, we can answer in the first iteration with parameter  $(d+1)(\varepsilon/2)$ , in the next with parameter  $(d+1)(\varepsilon/2)$ , in the next with parameter  $(d+1)(\varepsilon/4)$ , and so on, each time using up half of the remaining "privacy budget."

## 5. GENERATING SYNTHETIC DATA

The idea of creating a synthetic data set whose statistics closely mirror those of the original data set, but which preserves privacy of individuals, was proposed in the statistics community no later than 1993 [20]. The lower bounds on noise discussed at the end of Section 1.1 imply that no such data set can safely provide very accurate answers to too many weighted subset sum questions, motivating the interactive approach to private data analysis discussed herein. Intuitively, the advantage of the interactive approach is that only the questions actually asked receive responses.

Against this backdrop, the non-interactive case was revisited from a learning theory perspective, challenging the interpretation of the noise lower bounds as a limit on the number of queries that can be answered privately [3]. This work, described next, has excited interest in solutions yielding noise in the range  $[\omega(\sqrt{n}), o(n)]$ .

Let  $X$  be a universe of data items and  $\mathcal{C}$  be a *concept class* consisting of functions  $c : X \rightarrow \{0, 1\}$ . We say  $x \in X$  *satisfies* a concept  $c \in \mathcal{C}$  if and only if  $c(x) = 1$ . A concept class can be extremely general; for example, it might consist of all rectangles in the plane, or all Boolean circuits containing a given number of gates.

Given a sufficiently large database  $D \in X^n$ , it is possible to privately generate a synthetic database that maintains approximately correct fractional counts for *all* concepts in  $\mathcal{C}$  (there may be infinitely many!). That is, letting  $S$  denote the synthetic database produced, with high probability over the choices made by the privacy mechanism, for every concept  $c \in \mathcal{C}$ , the fraction of elements in  $S$  that satisfy  $c$  is approximately the same as the fraction of elements in  $D$  that satisfy  $c$ .

The minimal size of the input database depends on the quality of the approximation, the logarithm of the cardinality of the universe  $X$ , the privacy parameter  $\varepsilon$ , and the *Vapnik-Chervonenkis dimension* of the concept class  $\mathcal{C}$  (for finite  $|\mathcal{C}|$  this is at most  $\log_2 |\mathcal{C}|$ ). The synthetic dataset, chosen by the exponential mechanism, will be a set of  $m = O(\text{VCdim}(\mathcal{C})/\gamma^2)$  elements in  $X$  ( $\gamma$  governs the maximum permissible inaccuracy in the fractional count.) Letting  $D$

denote the input dataset and  $\hat{D}$  a candidate synthetic dataset, the utility function for the exponential mechanism is given by

$$u(D, \hat{D}) = -\max_{h \in \mathcal{C}} \left| h(D) - \frac{n}{m} h(\hat{D}) \right|.$$

## 6. PAN-PRIVACY

Data collected by a curator for a given purpose may be subject to "mission creep" and legal compulsion, such as a subpoena. Of course, we could analyze data and then throw it away, but can we do something even stronger, never storing the data in the first place? Can we strengthen our notion of privacy to capture the "never store" requirement?

These questions suggest an investigation of differentially private streaming algorithms with small state – much too small to store the data. However, nothing in the definition of a streaming algorithm, even one with very small state, precludes storing a few individual data points. Indeed, popular techniques from the streaming literature, such as Count-Min Sketch and subsampling, do precisely this. In such a situation, a subpoena or other intrusion into the local state will breach privacy.

A *pan-private* algorithm is private "inside and out," remaining differentially private even if its internal state becomes visible to an adversary [9]. To understand the pan-privacy guarantee, consider click stream data. These data are generated by individuals, and an individual may appear many times in the stream. Pan-privacy requires that any two streams differing only in the information of a single individual should produce very similar distributions on the *internal states* of the algorithm *and on its outputs*, even though the data of an individual are interleaved arbitrarily with other data in the stream.

As an example, consider the problem of *density estimation*. Assuming, for simplicity, that the data stream is just a sequence of IP addresses in a certain range, we wish to know what fraction of the set of IP addresses in the range actually appears in the stream. A solution inspired by randomized response can be designed using the following technique [9].

Define two probability distributions,  $D_0$  and  $D_1$ , on the set  $\{0, 1\}$ .  $D_0$  assigns equal mass to zero and to one.  $D_1$  has a slight bias towards 1; specifically, 1 has mass  $1/2 + \varepsilon/4$ , while 0 has mass  $1/2 - \varepsilon/4$ .

Let  $X$  denote the set of all possible IP addresses in the range of interest. The algorithm creates a table, with a one-bit entry  $b_x$  for each  $x \in X$ , initialized to an independent random draw from  $D_0$ . So initially the table is roughly half zeroes and half ones.

In an atomic step, the algorithm receives an element from the stream, changes state, and discards the element. When processing  $x \in X$ , the algorithm makes a fresh random draw from  $D_1$ , and stores the result in  $b_x$ . This is done no matter how many times  $x$  may have appeared in the past. Thus, for any  $x$  appearing at least once,  $b_x$  will be distributed according to  $D_1$ . However, if  $x$  never appears, then the entry for  $x$  is the bit drawn according to  $D_0$  during the initialization of the table.

As with randomized response, the density in  $X$  of the items in the stream can be approximated from the number of 1's in the table, taking into account the expected fraction of "false positives" from the initialization phase and the "false negatives" when sampling from  $D_1$ . Letting  $\theta$  denote the fraction of entries in the table with value 1, the output is

$4(\theta - 1/2)/\varepsilon + \text{Lap}(1/\varepsilon|X)$ .

Intuitively, the internal state is differentially private because, for each  $b \in \{0, 1\}$ ,  $e^{-\varepsilon} \leq \Pr_{\mathcal{D}_1}[b]/\Pr_{\mathcal{D}_0}[b] \leq e^\varepsilon$ ; privacy for the output is ensured by the addition of Laplacian noise. Over all, the algorithm is  $2\varepsilon$ -differentially pan-private.

## 7. CONCLUSIONS

The differential privacy frontier is expanding rapidly, and there is insufficient space here to list all the interesting directions currently under investigation by the community. We identify a few of these.

*The Geometry of Differential Privacy.* Sharper upper and lower bounds on noise required for achieving differential privacy against a sequence of linear queries can be obtained by understanding the geometry of the query sequence [13]. In some cases dependencies among the queries can be exploited by the curator to markedly improve the accuracy of the responses. Generalizing this investigation to the non-linear and interactive cases would be of significant interest.

*Algorithmic Complexity.* We have so far ignored questions of computational complexity. Many, but not all, of the techniques described here have efficient implementations. For example, there are instances of the synthetic data generation problem that, under standard cryptographic assumptions, have no polynomial time implementation [10]. It follows that there are cases in which the exponential mechanism has no efficient implementation. When can this powerful tool be implemented efficiently, and how?

*An Alternative to Differential Privacy?* Is there an alternative, “*ad omnia*,” guarantee that composes automatically, and permits even better accuracy than differential privacy? Can cryptography be helpful in this regard [19]?

The work described herein has, for the first time, placed private data analysis on a strong mathematical foundation. The literature connects differential privacy to decision theory, economics, robust statistics, geometry, additive combinatorics, cryptography, complexity theory learning theory, and machine learning. Differential privacy thrives because it is natural, it is not domain-specific, and it enjoys fruitful interplay with other fields. This flexibility gives hope for a principled approach to privacy in cases, like private data analysis, where traditional notions of cryptographic security are inappropriate or impracticable.

## 8. REFERENCES

- [1] N. R. Adam and J. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21:515–556, 1989.
- [2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The SuLQ framework. In *Proc. 24th ACM Symposium on Principles of Database Systems*, pages 128–138, 2005.
- [3] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proc. 40th ACM SIGACT Symposium on Theory of Computing*, pages 609–618, 2008.
- [4] D. E. Denning. Secure statistical databases with random sample queries. *ACM Transactions on Database Systems*, 5:291–315, 1980.
- [5] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proc. 22nd ACM Symposium on Principles of Database Systems*, pages 202–210, 2003.
- [6] C. Dwork. Differential privacy. In *Proc. 33rd International Colloquium on Automata, Languages and Programming (ICALP)(2)*, pages 1–12, 2006. See also: C. Dwork and M. Naor, On the difficulties of disclosure prevention in statistical databases, in submission.
- [7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. 3rd Theory of Cryptography Conference*, pages 265–284, 2006.
- [8] C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of lp decoding. In *Proc. 39th ACM Symposium on Theory of Computing*, pages pp. 85–94, 2007.
- [9] C. Dwork, M. Naor, T. Pitassi, G. Rothblum, and S. Yekhanin. Pan-private streaming algorithms. In *Proc. 1st Symposium on Innovations in Computer Science*, 2010.
- [10] C. Dwork, M. Naor, O. Reingold, G. Rothblum, and S. Vadhan. When and how can privacy-preserving data release be done efficiently? In *Proc. 41st International ACM Symposium on Theory of Computing*, pages 381–390, 2009.
- [11] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In *Advances in Cryptology – CRYPTO’04*, pages 528–544, 2004.
- [12] S. Goldwasser and S. Micali. Probabilistic encryption. *JCSS*, 28:270–299, 1984.
- [13] M. Hardt and K. Talwar. On the geometry of differential privacy. arXiv:0907.3754v2, 2009.
- [14] K. Kenthapadi, N. Mishra, and K. Nissim. Simulatable auditing. In *Proc. 24th ACM Symposium on Principles of Database Systems*, pages 118–127, 2005.
- [15] J. Kleinberg, C. Papadimitriou, and P. Raghavan. Auditing boolean attributes. In *Proc. 19th ACM Symposium on Principles of Database Systems*, pages 86–91, 2000.
- [16] R. Kumar, J. Novak, B. Pang, and A. Tomkins. On anonymizing query logs via token-based hashing. In *Proc. WWW 2007*, pages 629–638, 2007.
- [17] F. McSherry. Privacy integrated queries (codebase). available on Microsoft Research downloads website. See also pages 19-30, Proc. SIGMOD 2009.
- [18] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proc. 48th Annual Symposium on Foundations of Computer Science*, 2007.
- [19] I. Mironov, O. Pandey, O. Reingold, and S. Vadhan. Computational differential privacy. In *Advances in Cryptology – CRYPTO’09*, pages 126–142, 2009.
- [20] D. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:462–468, 1993.
- [21] L. Sweeney. Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics*, 25:98–110, 1997.
- [22] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *JASA*, pages 63–69, 1965.



WILLIAM DUKE/PHOTO BY DAXIAO PRODUCTIONS/SHUTTERSTOCK

Science's extensive COVID-19 coverage is free to all readers. To support our nonprofit science journalism, please **make a tax-deductible gift today.**

**Got a tip?**

**How to contact the news team**

Advertisement

## Can a set of equations keep U.S. census data private?

By **Jeffrey Mervis** | Jan. 4, 2019 , 2:50 PM

The U.S. Census Bureau is making waves among social scientists with what it calls a “sea change” in how it plans to safeguard the confidentiality of data it releases from the decennial census.

The agency announced in September 2018 that it will apply a mathematical concept called differential privacy to its release of 2020 census data after conducting experiments that suggest current approaches can't assure confidentiality. But critics of the new policy believe the Census Bureau is moving too quickly to fix a system that isn't broken. They also fear the changes will degrade the quality of the information used by thousands of researchers, businesses, and government agencies.

The move has implications that extend far beyond the research community. Proponents of differential privacy say a fierce, ongoing legal battle over plans to add a citizenship question to the 2020 census has only

underscored the need to assure people that the government will protect their privacy.

## A noisy conflict

The Census Bureau's job is to collect, analyze, and disseminate useful information about the U.S. population. And there's a lot of it: The agency generated some 7.8 billion statistics about the 308 million people counted in the 2010 census, for example.

At the same time, the bureau is prohibited by law from releasing any information for which "the data furnished by any particular establishment or individual ... can be identified."

Once upon a time, meeting that requirement meant simply removing the names and addresses of respondents. Over the past several decades, however, census officials have developed a bag of statistical tricks aimed at providing additional protection without undermining the quality of the data.

Such perturbations, also known as injecting noise, are meant to foil attempts to reidentify individuals by combining census data with other publicly available information, such as credit reports, voter registration rolls, and property records. But preventing reidentification has grown more challenging with the advent of ever-more-powerful computational tools capable of stripping away privacy.

Census officials now believe those ad hoc methods are no longer good enough to satisfy the law. "The problem is real, and it has moved from a concern to an issue," says John Thompson, who stepped down as census director in June 2017, and who recently retired as head of the Council of Professional Associations on Federal Statistics in Arlington, Virginia. "In Census Bureau lingo, that means it's no longer simply a risk, but rather something you have to deal with."

---

Advertisement

The agency's decision to adopt differential privacy was spurred, in part, by recent work on what is known as the "database reconstruction theorem." The theorem shows that, given access to a sufficiently large amount of information, someone can reconstruct underlying databases and, in theory, identify individuals.

"Database reconstruction theorem is the death knell for traditional [data] publication systems from confidential sources," says John Abowd, chief scientist and associate director for research at the Census Bureau, located in Suitland, Maryland. "It exposes a vulnerability that we were not designing our systems to address," says Abowd, who has spearheaded the agency's efforts to adopt differential privacy.

But some users of census data strongly disagree. Steven Ruggles, a population historian at the University of Minnesota in Minneapolis, is leading the charge against the new policy.

Ruggles says traditional methods have successfully prevented any identity disclosures and, thus, there's no urgency to do more. If the Census Bureau is hell-bent on imposing differential privacy, he adds, officials should work with the community to iron out the kinks before applying it to the 2020 census and its smaller cousin, the American Community Survey.

"Differential privacy goes above and beyond what is necessary to keep data safe under census law and precedent," says Ruggles, who also manages a university-based social research institute that disseminates census data. "This is not the time to impose arbitrary and burdensome new rules that will sharply restrict or eliminate access to the nation's core data sources."

"My central concern about differential privacy is that it's a blunt instrument," he adds. "If you want to provide the same level of protection against reidentification that current methods do, you're going to have to do a lot more damage to the data than is done now."

## Related Jobs

### Associate Director, Clinical Pharmacology

Pfizer  
La Jolla, California

### 2021 Global Talents Recruitment Announcement of Wuhan Textile University

Wuhan Textile University  
Wuhan, Hubei (CN)

### Sr. Research Assistant - Epigenetics and Molecular Carcinogenesis

University of Texas MD Anderson Cancer Center  
Houston, Texas

[MORE JOBS ►](#)

## Latest News

### Trending

- [1. The legendary dire wolf may not have been a wolf at all](#)
- [2. Japan plans to release Fukushima's contaminated water into the ocean](#)
- [3. 'Sink into your grief.' How one scientist confronts the emotional toll of climate change](#)
- [4. Chinese COVID-19 vaccine maintains protection in variant-plagued Brazil](#)
- [5. National academy may eject two famous scientists for sexual harassment](#)

### Most Read

- [1. Hard choices emerge as link between AstraZeneca vaccine and rare](#)
- [2. One number could help reveal how infectious a COVID-19 patient is.](#)



## Ways to protect confidentiality

Protecting confidentiality has been a priority for the Census Bureau for most—but not all—of its existence. After the first U.S. census was conducted in 1790, officials posted the results so that residents could correct errors. But in 1850, the interior secretary decreed that the returns would be kept confidential. They were “not to be used in any way to the gratification of curiosity and census officials,” or “the exposure of any man’s business or pursuits,” notes an official history of the census published in 1900. In 1954 the agency’s confidentiality mandate was codified in Title 13 of the U.S. Code.

Publicly available census data come in two flavors. One type, called small-area data, provides the basic characteristics of residents—age, sex, and race/ethnicity—down to the census block level. A census block, often the size of a city block, is the smallest geographic area for which data are reported. There were some 11 million blocks in 2010, of which 6.3 million were inhabited.

The second is called microdata, which are the full records collected by the Census Bureau on individuals—including, for example, the size of the household and the relationships between the residents. When microdata are reported, they are lumped together by areas containing at least 100,000 people.

Together, these census products provide fodder for thousands of researchers. Census data are also the basis for surveys by other government agencies and the private sector that shape decisions ranging from locating new factories or shopping malls to building new roads and schools.

The Census Bureau has used a variety of methods to preserve the confidentiality of these data as it moved from print to magnetic tape to digital distribution. Officials can, for instance, mask the responses of outliers—such as the income of a billionaire. They can also be less precise, for example, by reporting ages within 5-year ranges rather than a single year. Another technique

**3. Top German psychologist fabricated data, investigation finds**

**4. Particle mystery deepens, as physicists confirm that the muon is more magnetic than predicted**

**5. Food supplements that alter gut bacteria could ‘cure’ malnutrition**

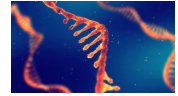
### Sifter

**This homely mollusk’s rock-hard chompers are made of rare minerals**



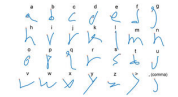
By Sofia Moutinho | Jun. 2, 2021

**Sugars spice up RNA**



By Robert Service | May. 18, 2021

**Paralyzed person types at record speed—by imagining handwriting**



By Kelly Servick | May. 13, 2021

**Climate change is triggering more lightning strikes in the Arctic**



By Sofia Moutinho | Apr. 7, 2021

**Tropical forest destruction increases, despite the pandemic**



By Sofia Moutinho | Mar. 31, 2021

**More Sifter**

involves swapping information with a respondent possessing many similar characteristics who lives in a different block.

How much noise to inject depends on many factors. However, census officials have never disclosed details of their formula or said how often a particular method is used. They fear that such information could help someone to reverse engineer the process.

## **A mathematical approach**

Differential privacy, first described in 2006, isn't a substitute for swapping and other ways to perturb the data. Rather, it allows someone—in this case, the Census Bureau—to measure the likelihood that enough information will “leak” from a public data set to open the door to reconstruction.

“Any time you release a statistic, you’re leaking something,” explains Jerry Reiter, a professor of statistics at Duke University in Durham, North Carolina, who has worked on differential privacy as a consultant with the Census Bureau. “The only way to absolutely ensure confidentiality is to release no data. So the question is, how much risk is OK? Differential privacy allows you to put a boundary” on that risk.

A database can be considered differentially protected if the information it yields about someone doesn't depend on whether that person is part of the database.

Differential privacy was originally designed to apply to situations in which outsiders make a series of queries to extract information from a database. In that scenario, each query consumes a little bit of what the experts call a “privacy budget.” After that budget is exhausted, queries are halted in order to prevent database reconstruction.

In the case of census data, however, the agency has already decided what information it will release, and the number of queries is unlimited. So its challenge is to

calculate how much the data must be perturbed to prevent reconstruction.

Abowd says the privacy budget “can be set at wherever the agency thinks is appropriate.” A low budget increases privacy with a corresponding loss of accuracy, whereas a high budget reveals more information with less protection. The mathematical parameter is called epsilon; Reiter likens setting epsilon to “turning a knob.” And epsilon can be fine-tuned: Data deemed especially sensitive can receive more protection.

The epsilon can be made public, along with the supporting equations on how it was calculated. In contrast, Abowd says, traditional approaches to limiting disclosure are “fundamentally dishonest” from a scientific perspective because of their underlying uncertainty. “At the moment,” he says, the public doesn’t “know the global disclosure risk. ... That’s because the agency doesn’t tell you everything it did to the data before releasing it.”

## **A simulated attack**

A professor of labor economics at Cornell University, Abowd first learned that traditional procedures to limit disclosure were vulnerable—and that algorithms existed to quantify the risk—at a 2005 conference on privacy attended mainly by cryptographers and computer scientists. “We were speaking different languages, and there was no Rosetta Stone,” he says.

He took on the challenge of finding common ground. In 2008, building on a long relationship with the Census Bureau, he and a team at Cornell created the first application of differential privacy to a census product. It is a web-based tool, called OnTheMap, that shows where people work and live.

Abowd took leave from Cornell to join the Census Bureau in June 2016, and one of his first moves was to test the vulnerability of the 2010 census data to an outside attack. The goal was to see how well a census team could

reconstruct individual records from the thousands of tables the agency had published—and then try to identify those individuals.

The three-step process required substantial computing power. First, the researchers reconstructed records for individuals—say, a 55-year-old Hispanic woman—by mining the aggregated census tables. Then, they tried to match the reconstructed individuals to even more detailed census block records (that still lacked names or addresses); they found “putative matches” about half the time.

Finally, they compared the putative matches to commercially available credit databases in hopes of attaching a name to a particular record. Even if they could, however, the team didn’t know whether they had actually found the right person.

Abowd won’t say what proportion of the putative matches appeared to be correct. (He says a forthcoming paper will contain the ratio, which he calls “the amount of uncertainty an attacker would have once they claim to have reidentified a person from the public data.”) Although one of Abowd’s recent papers notes that “the risk of re-identification is small,” he believes the experiment proved reidentification “can be done.” And that, he says, “is a strong motivation for moving to differential privacy.”

## Too far, too fast?

Such arguments haven’t convinced Ruggles and other social scientists opposed to applying differential privacy on the 2020 census. They are circulating manuscripts that question the significance of the census reconstruction exercise and that call on the agency to delay and change its plan.

Last month they had their first public opportunity to express their opposition during a meeting at census headquarters of the Federal Economic Statistics Advisory

Committee (FESAC), which advises the Census Bureau and two other major federal statistical agencies. Abowd and Ruggles went toe to toe during a panel discussion on differential privacy, and council members had a chance to quiz them.

One point of disagreement is the interpretation of federal law. Title 13 requires the agency to mask only the identity of individuals, critics argue, not their characteristics. If identifying characteristics is illegal, Ruggles writes in a recent paper, then “virtually all Census Bureau microdata and small-area products currently fail to meet that standard.”

Abowd reads the law differently. “Steve has gotten it wrong,” he says flatly. “The statute says that what is prohibited is releasing the data in an identifiable way.”

At the meeting, several members of the advisory committee peppered Abowd with questions about the significance of being able to reconstruct 50% of microdata files. That percentage is rather low, they argue. In any event, they say, reconstruction is a far cry from reidentification, which is what the law prohibits. They also wondered why anyone would go to the trouble of messing with census data when there are other, better ways to obtain scads of personal information that can be used to identify individuals.

“I’m not surprised that someone has reconstructed the fact that there are 45-year-old white men living in a particular block,” said Colm O’Muircheartaigh, a professor of public policy at the University of Chicago in Illinois and a member of FESAC. “But that kind of information is neither very interesting or useful.”

Identifying individuals based on household data might be more valuable, he said. “But I imagine it would be much harder to reconstruct a household,” O’Muircheartaigh said. “And even if we could, reconstructing a typical American household—say, two adults and two children—would hardly be a killer identification.”

Census data also don't age well because of high mobility rates, he added. "These are static data," he said. "Even if you knew that such and such a person lived somewhere in 2010, how valuable would that be in 2014 or 2018?"

Some meeting attendees also accused Abowd of failing to address the practical effects of applying differential privacy. One skeptic was Kirk Wolter, chief statistician for NORC at the University of Chicago, a research institution that does survey work for many federal agencies. He argued that noisier census data would have a major ripple effect, degrading the quality of many other surveys that rely on census data to select their samples. "These surveys provide the information infrastructure for the country," he noted. "And all of them would suffer."

Correcting for those problems will cost money, he predicted, with organizations like NORC having to adjust samples and redesign surveys. And given the tight budgets of most survey research organizations, those could translate into fewer studies—and less information about the country's residents.

Thompson agrees. "Kirk is exactly right," he says. Applying differential privacy means "those surveys will take longer and cost more. And they may be less accurate. But you don't have a choice."

## The citizenship elephant

Proponents of adopting differential privacy say there is also another compelling reason to move forward quickly: a controversial decision made last March by Commerce Secretary Wilbur Ross to **add a citizenship question** to the 2020 census.

A slew of local and state officials have joined civil rights groups in suing the federal government in a bid to block the question. They argue that adding the question will lead nonresidents and other vulnerable populations to avoid filling out the census form, **leading to a significant undercount**. And they are worried about privacy, too.

Knowing how someone answered the citizenship question, critics say, would allow a government agency to take punitive action against nonresidents.

“Maybe a researcher wouldn’t try to do that,” says Thompson, a witness for the plaintiffs in one of the suits. “But there are a lot of people who might. And I think that [federal immigration officials] would love to have that information.”

Abowd knows the extreme sensitivity of the citizenship question. His emails last year to Ross expressing reservations about adding it to the 2020 census have been publicly revealed by the litigation. And although he tiptoed around the topic during the recent FESAC discussion, it was clear that he was worried about the damage it could wreak on the agency’s credibility.

“The entire history of traditional disclosure limitation was aimed at preventing attackers, armed with external data, from using it in combination with the variables on the [census] microdata file to attach a name and address,” Abowd said during the roundtable. “With regard to 2010, most of those databases did not have race and ethnicity on them. And none have citizenship, to just bring into the room the variable that we probably should be discussing more explicitly.”

## Practical issues

Ruggles, meanwhile, has spent a lot of time thinking about the kinds of problems differential privacy might create. His Minnesota institute, for instance, disseminates data from the Census Bureau and 105 other national statistical agencies to 176,000 users. And he fears differential privacy will put a serious crimp in that flow of information.

In the most extreme scenario, he says, the Census Bureau could decide to make 2020 census data available only through its network of 29 secure Federal Statistical Research Data Centers. That would impose serious

hardships on users, Ruggles says, because the centers require users to obtain a security clearance, which often involves lengthy waiting periods. Such rules could also prevent most international scholars from using the centers, he says, as well as graduate students seeking a quick turnaround for a dissertation. In addition, researchers are only cleared if their project is deemed to benefit the agency's mission.

There are also questions of capacity and accessibility. The centers require users to do all their work onsite, so researchers would have to travel, and the centers offer fewer than 300 workstations in total.

Thompson says the Census Bureau needs to address those issues regardless of whether it adopts differential privacy. He agrees with Ruggles that it takes too long to gain access to the research centers, and he thinks the bureau needs to change its definition of what research serves its mission. "I have argued that anyone advancing the science of using data" should be eligible, he says. "We need a 21st-century Census Bureau, and that will take a lot of fixing."

(With regard to access, Abowd says the agency is considering setting up "virtual" centers that would allow a much broader audience to work with the data. But Ruggles is skeptical that such a system would satisfy the bureau's own definition of confidentiality.)

## **A need to communicate**

Abowd has said, "The deployment of differential privacy within the Census Bureau marks a sea change for the way that official statistics are produced and published." And Ruggles agrees. But he says the agency hasn't done enough to equip researchers with the maps and tools needed to navigate the uncharted waters.

"It's pretty clear we are going to have a new methodology," Ruggles concedes. "But I think it could be implemented in a better or worse way. I would like them



to consider the trade-offs, and not take such an absolutist stand on the risks.”

Meanwhile, NORC’s Wolter says regardless of whether his concerns are addressed, the bureau must do more outreach—and not just in peer-reviewed journals. “Census badly needs a communications strategy, by real communications specialists,” he said. “There are thousands of users [of census data] who won’t understand any of this stuff. And they need to know what is going to happen.”

*Clarification, 17 January 2019, 5:00 p.m.: The first quote from John Abowd in the story has been revised to make it clear that the Census Bureau is now addressing the vulnerability of census data to reidentification.*

Posted in: [Science and Policy](#), [Scientific Community](#)  
doi:10.1126/science.aaw5470



**Jeffrey Mervis**

Jeff tries to explain how government works to readers of *Science*.

[Email Jeffrey](#)

## More from News

**NIH should boost rigor of animal studies with stronger statistics, pilot studies, experts say**



**Europe loosens funding rules for non-EU quantum and space researchers**

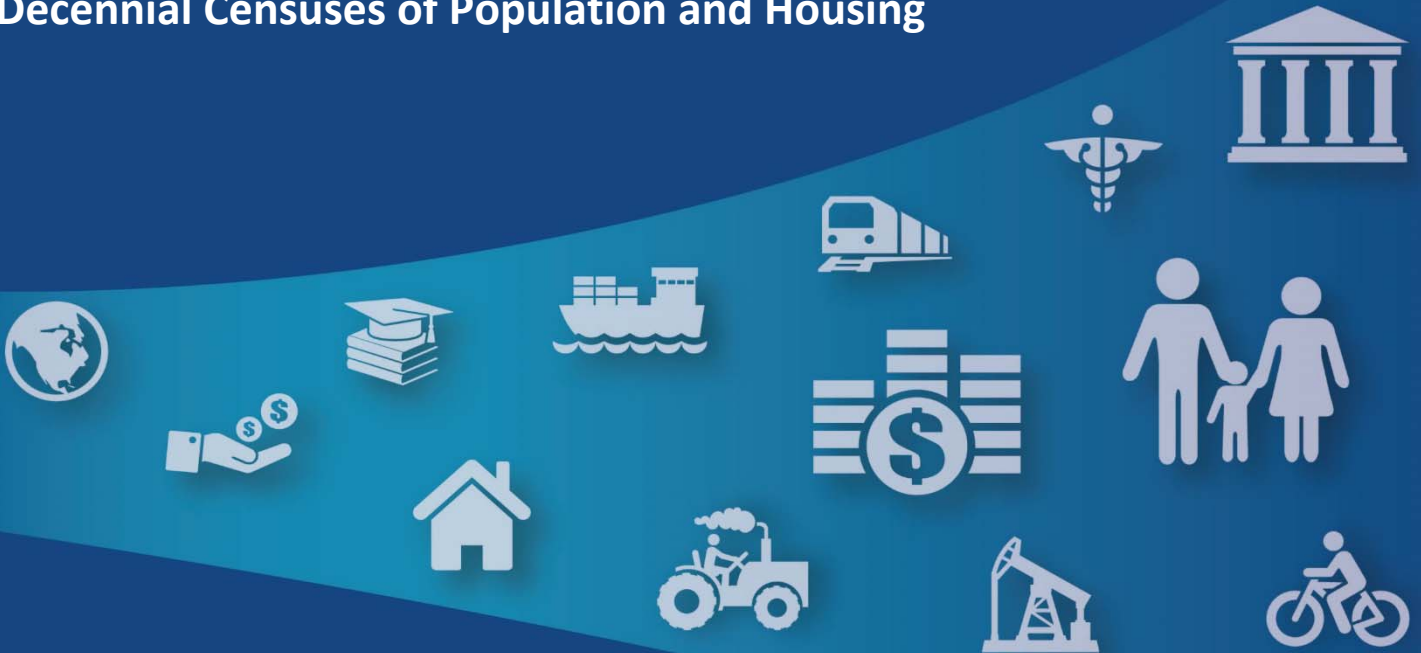


**House science panel firms up its plan to expand NSF**



# THE RESEARCH AND METHODOLOGY DIRECTORATE

## Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing



Laura McKenna

November 2018



# Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing

Laura McKenna<sup>1</sup>

October 2018

The U.S. Census Bureau conducts the decennial censuses under Title 13 of the U. S. Code with the Section 9 mandate to not “use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports (13 U.S.C. § 9 (2007)).” The Census Bureau applies *disclosure avoidance* techniques to its publicly released statistical products in order to protect the confidentiality of its respondents and their data.

## Foreword

John M. Abowd

Chief Scientist and Associate Director for Research and Methodology

Laura McKenna is the former Chief of the Center for Disclosure Avoidance Research and former Chair of the Disclosure Review Board. I asked her to write this overview of the disclosure avoidance methods used in the last five decennial censuses in order to guide contemporary readers through that history in single document and with a coherent vocabulary. In September 2017, the Census Bureau announced that it would undertake a comprehensive disclosure avoidance modernization program beginning with the 2020 Census of Population and Housing. The 2020 census will be protected by modern formal privacy methods—specifically, differential privacy, continuing the long history of innovation in confidentiality protection documented in this review.

---

<sup>1</sup> This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau. Thanks to Connie Citro, Cynthia Clark, Jerry Gates, Nancy Gordon, Michele Hedrick, Bud Pautler, and Sara Sullivan for background, and other assistance in preparing this report.

## Contents

1	Introduction .....	1
2	1970 Census of Population and Housing .....	3
3	1980 Census of Population and Housing .....	4
3.1	Why change the methods from the 1970 Census? .....	4
3.2	100% Data (PL 94-171, Summary File (SF) 1, and SF2) .....	4
3.3	Sample Data (SF3 and SF4) .....	5
3.4	Household Data .....	5
4	1990 Census of Population and Housing .....	6
4.1	Why change the methods from the 1980 Census? .....	6
4.2	100% Data (PL 94-171, SF1, and SF2) .....	6
4.3	Sample Data (SF3 and SF4) .....	7
4.4	Household Data .....	7
5	2000 Census of Population and Housing .....	8
5.1	Why change the methods from the 1990 Census? .....	8
5.2	100% Data (PL 94-171, SF1, and SF2) .....	8
5.3	Sample Data (SF3 and SF4) .....	8
5.4	Household Data .....	9
6	2010 Census of Population and Housing .....	10
6.1	Why change the methods from the 2000 Census? .....	10
6.2	100% Data (PL 94-171, SF1, and SF2) .....	10
6.3	Sample Data (SF3 and SF4) .....	10
6.4	Household Data .....	10
6.5	Group Quarters Data .....	10
7	Conclusion .....	11
8	References .....	12
9	Appendix A .....	13
10	Appendix B.....	15
11	Appendix C.....	30
12	Appendix D .....	32
13	Appendix E.....	34

## 1 Introduction

The U.S. Census Bureau’s disclosure avoidance (DA) methods have evolved over the past 50 years. The 2020 Census will be the first census protected by a formally private disclosure avoidance system based on differentially privacy methods. This paper summarizes the historical methods the agency used from the 1970 to the 2010 censuses, leading up to the adoption of the modernized disclosure avoidance methods.

This history discusses only publicly available information about the confidentiality protection methods as noted in official documentation of the relevant decennial censuses. All of the information in this summary was taken from historical public sources, except as noted. None of the information in this paper is confidential.

There is no public documentation of the disclosure avoidance methods used in the 1970 Census. This paper relies on an internal Census Bureau planning paper, now cleared for release, that provided a brief description of 1970 methods while highlighting options for disclosure avoidance for the 1980 Census (Zeisset, 1978). There is no information about 1970 methods in the 1970 Technical Documentation nor the 1970 Data User’s Guide. Likewise, no documentation of disclosure avoidance was found in public or internal papers for pre-1970 censuses.

The first documented discussion of disclosure avoidance techniques for Group Quarters (GQ) data was for the 2010 Census. There is no discussion of disclosure avoidance for GQ data in public or internal documents for the 1980, 1990, and 2000 censuses.<sup>2</sup>

This paper is focused on decennial census tabular data. A separate paper will outline the history of disclosure avoidance methodology for the Public Use Microdata Samples (PUMS) files. The American Community Survey (ACS) is out of scope for both papers.

This history gleans procedures from various types of publications (Public Law 94-171, Summary Files 1-4) and for different tabulation populations—people in households, people in Group Quarters, 100% (“short form”) data, and sample (“long form”) data. Complete enumeration (100% data) is used for Public Law (PL) 94-171 (data for redistricting purposes), Summary File (SF) 1, and SF2. Through the 2000 Census, sample data were published in SF3 and SF4. The 2010 Census was the first recent census not to include long form data; the ongoing ACS replaced that data source starting in 2005. All publications were based on both people in households and people in Group Quarters. Tables in SF2 were similar to tables in SF1, but they were iterated by race and Hispanic origin. Tables in SF4 were similar to tables in SF3, but they were iterated by race and Hispanic origin.

---

<sup>2</sup> Group Quarters data include information about people living in nursing homes, prisons, college dormitories, military barracks, etc. (somewhere other than a household).

The Census Bureau did not publish long form sample data at the lowest level of geography (blocks). As is still the practice with the long form's successor, the American Community Survey, the smallest published geography is the block *group* level.

Rules for special tabulations from the 2000 and 2010 decennial censuses (Appendix A) added another layer of confidentiality protection by restricting releasable special tabulation details.

Notes on Confidentiality in the Technical Documentation of the 1980 through 2010 censuses (Appendices B, C, D, and E) provided high-level information about confidentiality protection in the decennial censuses. Census Bureau researchers published additional details about methods through working papers and symposia and continue to do so. Today, data users can request information or ask questions by contacting disclosure avoidance subject matter experts at [DRB\\_CHAIR@census.gov](mailto:DRB_CHAIR@census.gov).

## 2 1970 Census of Population and Housing

The Census Bureau relied on whole table suppression — not individual cell suppression — as the primary disclosure avoidance method for the 1970 Census. Table suppression was based on the number of people or households in a given area. The method was problematic for several reasons:

1. fewer tables were available for data users;
2. the agency did not provide guidance on how to account for the suppressed data when analyzing the published data;
3. the protections brought by the suppressed whole tables were diminished by the fact that very few complementary tables were suppressed; and
4. cells within an original table could still show an original estimate of 1 or 2.

To limit disclosure risk, the lowest geographic level for which sample data were published was (and still is) the census block group. Census 100% data were published for the lowest possible geographic level: census blocks.

All disclosure avoidance information from the 1970 Census was obtained from an internal document (Zeisset, 1978).



### 3 1980 Census of Population and Housing

#### 3.1 Why change the methods from the 1970 Census?

Data user dissatisfaction with whole table suppression, along with concerns about the lack of complementary table suppression, lead the Census Bureau to explore new disclosure avoidance methods for the 1980 Census. Researchers discussed options that included random rounding, ordinary rounding, combining areas, and table redesign (Zeisset, 1978). Ultimately, the Census Bureau chose to continue using table suppression, but added additional suppression of complementary tables.

#### 3.2 100% Data (PL 94-171, Summary File (SF) 1, and SF2)

The agency used table-level data suppression for 1980 census tabular data products (Griffin et al., 1989). As in 1970, some tables with cell estimates of 1 or 2 were published. In this case, the counts were replaced with 0s and a flag designating that the cell was suppressed for disclosure, but complementary suppressions were not applied (see Appendix B).

The following univariate (one-variable) counts were not suppressed at any geographic level, the smallest being the block level (for 100% data):

- Population counts by race or Hispanic origin.
- Housing unit counts by vacancy status.
- Occupied housing unit counts by race or Hispanic origin of the householder.

The following rules were applied to data for blocks and above (larger geographical areas) (100% data) and for block groups and above (sample data). A suppression universe is defined as one variable or the cross tabulation of a very small set of variables for which many tables are iterated, such as was the case in SF2 and SF4 (which iterate SF1 and SF3, respectively, across multiple race and Hispanic origin categories).

- **Race or Hispanic origin of householder:**
  - **1 to 14 people:** Detailed characteristics collected for total population, or any suppression universe defined by race or Hispanic origin of the householder, were suppressed if there were 1 to 14 people in the specified suppression universe (for example Black female householders in a given geographic area).
  - **1 to 4 occupied housing units:** Detailed characteristics for people in households for suppression universes defined by the race or Hispanic origin of the householder were suppressed if there were 1 to 4 occupied housing units in the specified group (for example White male householders who rent in a given geographic area).
- **Vacancy status:**
  - **1 to 4 vacant and or occupied housing units:** Detailed housing characteristics for suppression universes defined by vacancy status were suppressed if there were 1

to 4 housing units in the relevant universe (for example Occupied housing units with running water in a given geographic area).

- **Complementary suppression:**

- **Race and tenure:** Complementary table suppression was applied to protect the additive relationships for race groups that added to a total and for tenure (owners + renters = total) in non-univariate iterated tables. Pre-established rules governed the sequence of choosing complementary table suppressions, for example, suppressing smallest to largest populated tables in a given area.
- **Cross-geographic areas:** A shortcoming of the 1980 methods was that complementary table suppression was not applied across geographic areas (Griffin et al., 1989). So, for example, if data for one of the three Delaware counties was suppressed, someone could uncover the suppressed tables by subtracting the data for the other two counties from data for the whole state.

### 3.3 Sample Data (SF3 and SF4)

See Section 3.2 which describes the method for 100% data and was also used for sample data.

### 3.4 Household Data

See Section 3.2 which describes the method for 100% data including households.

## 4 1990 Census of Population and Housing

### 4.1 Why change the methods from the 1980 Census?

Census Bureau researchers developed new disclosure avoidance methods to address three primary shortcomings of 1980 methods:

- dissatisfaction with the reduction in data tables caused by whole table suppression;
- the lack of guidance for data users using the published data in the presence of suppression;
- the disclosure risk issues caused by the lack of complementary suppression across geographic areas (Griffin et al., 1989).

### 4.2 100% Data (PL 94-171, SF1, and SF2)

Data were published at all geographic levels, including the smallest level, blocks.

The Census Bureau replaced whole table suppression with a new disclosure avoidance technique for the 1990 Census. The new “Confidentiality Edit” used rules-based “data swapping” at the microdata (individual record) level (known then as the “data interchange” method) for 100% data, and the “Blank and Impute” technique for sample data (see Section 4.3).

For 100% data it kept the following unchanged:

- population counts by total, race, Hispanic origin, and people of age 18 and above;
- housing unit counts by total, tenure, and rent/value categories.

To apply the Confidentiality Edit, agency data staff:

1. Selected a small sample of households from the internal census data files, with a higher sampling rate for small blocks.
2. Paired the sampled records according to a set of well-defined matching rules to other records on the file in different geographic locations.
3. Maintained a 1-to-1 matching basis for key variables between each sampled household and its paired household in the other geographic location for the following variables:
  - household size;
  - householder race;
  - householder Hispanic origin;
  - number of people age (18+);
  - tenure (own/rent); and
  - rent/value category.
4. “Interchanged” the paired household records according to a well-defined data interchange (data swapping) operation. The “interchanged” file (swapped file) became the official version of the internal detail file and was used to prepare all subsequent

census data products. A brief discussion of the evaluation of this method is available (Griffin et al., 1989).

#### 4.3 Sample Data (SF3 and SF4)

For all published areas except small block groups, the fact that data were data from a sample was judged to provide adequate disclosure protection.

For small block groups, Census researchers developed what became known as the “Blank and Impute” technique. It involved “blanking” (removing) a sample of the data values (population and housing items) for one of the sample housing units in each small block group and imputing those values using the 1990 Census imputation methodology.

The resulting sample data file (to which disclosure avoidance had been applied) was used to prepare all subsequent census sample data products.

Primarily because of the relatively small increase in imputation rates, the Blank and Impute technique added very little to the level of error of the estimates (Griffin et al., 1989).

#### 4.4 Household Data

The techniques described in Sections 4.2 and 4.3 were used for household data.

## 5 2000 Census of Population and Housing

### 5.1 Why change the methods from the 1990 Census?

For the 1990 Census, Census Bureau researchers applied new DA techniques targeted to one of the riskiest potential disclosure categories: small blocks and block groups. For the 2000 Census, staff sought to extend these types of protections beyond small geographies to other increased-risk categories, particularly those at greater risk due to unique cross-tabulations and key variables.

The 2000 Census was the first to allow respondents to choose multiple race categories. The additional detail brought with it a new total of 63 possible race “alone” or “combined” answers. This posed a significant disclosure avoidance challenge and prompted the Census to apply additional protections.

After the 1990 Census the science of disclosure avoidance continued to evolve, and the Census Bureau extended swapping-based protections to the 2000 Census. Swapping replaced Blank and Impute as the primary disclosure protection method for sample data. Swapping had the advantage of removing any absolute assurance that a given record belonged to a given household. It also retained relationships among the variables for each household.

### 5.2 100% Data (PL 94-171, SF1, and SF2)

Census researchers expanded the swapping techniques inaugurated in 1990 to additional higher-risk categories for the 2000 Census as follows:

- The probability of swapping increased for cross-tabulations of key variables, smaller blocks, and for households that contained members of a race category not found in other households in that block.
- The probability of swapping decreased for blocks already protected with high imputation rates. Records that were entirely imputed were not swapped.
- Every record not totally imputed had a small chance of being swapped.
- Pairs of households that were swapped matched on a second set of key demographic variables. All data products were created from the swapped file.
- For the SF2 dataset, a minimum of 100 people of a race or Hispanic origin group (Hispanic/Non-Hispanic) were required in a geographic area to publish a table iterated by that group for that area. (Zayatz, 2003; Zayatz, 2007). No complementary suppression was applied in order to preserve data quality and save paper.

### 5.3 Sample Data (SF3 and SF4)

The same disclosure avoidance methods were applied to the sample data at the block group level, with the following differences:

- In addition to decreased swapping rates for block groups with higher imputation rates, rates also decreased in block groups with lower sampling rates.
- For the SF4 dataset, a minimum of 50 people of a race or Hispanic origin group were required in a geographic area to publish a table iterated by that group for that area. (Zayatz, 2003; Zayatz, 2007). No complementary suppression was applied in order to preserve data quality and save paper.
- Sample data required a third list of variables to be held fixed (unswapped). For example, some variables between paired households weren't swapped, such as a householder's American Indian tribe. All three of the lists of variables are confidential.

#### 5.4 Household Data

The household data were protected using data swapping as described in Sections 5.2 and 5.3.

## 6 2010 Census of Population and Housing

### 6.1 Why change the methods from the 2000 Census?

The 2010 Census was the first “short form-only” census in recent history. The former sample data long form was replaced by the ongoing “American Community Survey.”

### 6.2 100% Data (PL 94-171, SF1, and SF2)

See Sections 6.4 and 6.5 below.

### 6.3 Sample Data (SF3 and SF4)

The 2010 Census did not include a long form. The questions previously asked on the long form were transferred to the new American Community Survey.

### 6.4 Household Data

The swapping procedures for household data were essentially the same as those used for Census 2000 with some refinements to the key variables used to identify unique records and the key variables used to find swapping partners (Zayatz et al., 2010).

### 6.5 Group Quarters Data

The Census Bureau developed Partially Synthetic Data models to protect Group Quarters (GQ) data. The process involved:

- Blanking some values in at-risk respondent records and using synthetic data techniques to impute those values.
- Using key variable cross tabulation to locate unique records in each tract.
- Blanking unique variable values within each record (compared to other records in the tract).
- Replacing the blanked values with predicted values developed from two types of generalized linear models developed for each county: polytomous regression models and generalized additive models. Variable values were processed in a specific order. Once a value was synthesized, it was used as a predictor for synthesizing other variables.
- Geography and type of GQ were never altered, and age groups <18, 18+ were held fixed.

## 7 Conclusion

The Census Bureau’s disclosure avoidance techniques have evolved over the decades. In 1970 and 1980, the agency used table suppression. Beginning with the 1990 Census, the agency used newer methods, applied at the microdata (individual record) level. In 1990, the “Confidentiality Edit” applied data swapping for 100% (short form) data and blanking and imputation for sample (long form) data.

Beginning in 2000, the Census Bureau extended data swapping to the sample data. While the actual swapping rate and its impact on overall accuracy is confidential, a confidential research study found that the impact in terms of introducing error into the estimates was much smaller than errors from sampling, non-response, editing and imputation.

In 2010, the agency generated partially synthetic data to protect Group Quarters data.

Throughout the decades, the agency published 100% data at the block level and above, and sample data at the block group level and above.

		Table Suppression	Swapping	Blank and Impute	Partially Synthetic Data
1970					
	100% Data	X			
	Sample Data	X			
	Households	X			
1980					
	100% Data	X			
	Sample Data	X			
	Households	X			
1990					
	100% Data		X		
	Sample Data			X	
	Households		X	X	
2000					
	100% Data		X		
	Sample Data		X		
	Households		X		
2010					
	100% Data		X		X
	Households		X		
	Group Quarters				X



## 8 References

Lauger, A., Wisniewski, W., and McKenna, L. (2015), "Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research," Proceedings of the Section on Government Statistics, American Statistical Association, pp. 3630-3642.

Griffin, R., Navarro, F., and Flores-Baez, L. (1989), "Disclosure Avoidance for the 1990 Census," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 516-521.

United States, Congress, House. United States Code. United States Government Printing Office, 30 Mar. 2010, <https://www.gpo.gov/fdsys/pkg/USCODE-2007-title13/pdf/USCODE-2007-title13.pdf>.

Zayatz, L. (2002), "SDC in the 2000 U.S. Decennial Census," In: Domingo-Ferrer, J. (eds) Inference Control in Statistical Databases. Lecture Notes in Computer Science, vol 2316, Springer, Berlin, Heidelberg.

Zayatz, L. (2003), "Disclosure Limitation for Census 2000 Tabular Data," Working Paper #15, ECE/Eurostat workshop on statistical data confidentiality, <http://www.unece.org/stats/documents/2003/04/confidentiality/wp.15.e.pdf>.

Zayatz, L. (2007), "Disclosure Avoidance Practices and Research at the US Census Bureau: An Update," Journal of Official Statistics 23 no. 2: pp. 253-265.

Zayatz, L., Lucero, J., Massell, P., Ramanayake, A. (2010), "Disclosure Avoidance for Census 2010 and American Community Survey Five-year Tabular Data Products," Section on Survey Research Methods, American Statistical Association, pp. 2279-2288.

Zeisset, P. (1978), "Suppression vs. Random Rounding: Disclosure Avoidance Alternatives for the 1980 Census," <https://www.census.gov/library/working-papers/series/cdar-wp.1978.html>

## 9 Appendix A

### Rules for special tabulations from the 2000 and the 2010 Decennial Censuses

1. All Decennial Census special tabulations must be reviewed by the Disclosure Review Board.
2. All cells in any special tabulation must be rounded. The rounding schematic is:

0 remains 0

1-7 rounds to 4

8 or greater rounds to nearest multiple of 5 (i.e., 864 rounds to 865, 982 rounds to 980)

Any number that already ends in 5 or 0 stays as is.

This rounding applies to all special tabulations that pertain to the population in households or the population in group quarters -- those done under reimbursable agreement, those done for working papers, tables, professional papers, etc.

Any totals or subtotals needed should be constructed before rounding. This assures that universes remain the same from table to table, and it is recognized that cells in a table will no longer be additive after rounding.

3. Medians or other quantiles may be calculated as
  - A. an interpolation from a frequency distribution of unrounded data (these are not subject to additional rounding), or
  - B. as a point quantile. These must be rounded to two significant digits: 12,345 would round to 12,000; 167,452 would round to 170,000. There must be at least 5 cases on either side of the quantile point. It is recognized that the interpolated quantile may indeed be some individual's response, but it is coincidental, not by design.
4. Tables for sample data are only published after weights have been applied to the data, but sometimes both weighted and unweighted counts are used when applying disclosure avoidance rules. Thresholds on universes will normally be applied to avoid showing data for very small geographic areas or for very small population groups, often 50 unweighted cases for sample data. Tables may normally not have more than 3 or 4 dimensions, and mean cell size lower limits may also be required. For example, the mean cell size of each table must be at least 3 cases for 100% data, or 20 weighted cases for sample data).
5. Percents, rates, etc., should be calculated after rounding, but the DRB has granted exceptions to this rule when the numerator and/or denominator of the percent or rate is not shown.
6. Means and aggregates must be based on at least 3 values.
7. The finest level of detail shown for Group Quarters data will be Institutional/ Noninstitutional.

8. For Demographic Profiles from user-defined geographic areas (neighborhoods), all areas must have at least 300 people in them. Using a computer program, the user-defined areas will be compared with standard Census Bureau areas to make sure users cannot obtain data from very small geographic areas by subtraction. If such small areas are found, the boundaries of the user-defined areas must be changed.

# APPENDIX B

## 1980, part 1

1980 Census of Population and  
Housing

Summary Tape File 2.

Technical Documentation

pp. 25-31

prepared by

the U.S. Census Bureau, 1982

"PRECEDING PAGE BLANK -- NOT FILMED"  
SUPPRESSION

To maintain the confidentiality promised respondents and required by law, the Census Bureau takes precautions to make sure that its public data, in print or on tape, do not disclose information about particular individuals or housing units; therefore, the Bureau suppresses tabulations of characteristics for very small groups of people or housing units. On summary tapes, zeroes are entered in suppressed cells. Flag fields which indicate suppression are shown on each record. However, a zero in a cell does not automatically mean suppression. Only by checking the suppression flag can it be determined if the zero in a specific table is suppressed data or an actual count of zero.

This discussion outlines the rules for suppression, how its occurrence can be identified by the user, and how to handle it.

No Suppression

The following counts will never be suppressed:

- Total population
- Total housing units
- Seasonal/migratory housing units
- Year-round housing units
- Occupied housing units
- Vacant year-round housing units
- Count of persons and households in each race and Spanish origin group

Primary Suppression

Suppression of Population Characteristics. Characteristics of persons other than race or Spanish origin (e.g., age, relationship) are shown only if there are 15 or more persons in the geographic area. For example, on a record for a census tract with a population of 1 to 14 persons, population characteristics such as age and relationship are suppressed. Only counts for total population and the number of persons within specific race or Spanish origin groups are provided.

However, when the geographic area being summarized has 15 or more persons, no suppression of population characteristics will occur--except possibly when tables are cross-classified by race or Spanish origin. The rules for this type of suppression are outlined below in Suppression of Tables Cross-Classified by Race or Spanish Origin.

Suppression of Year-round Housing Characteristics. Characteristics of year round housing units which are not classified by occupancy status (e.g., number of rooms, plumbing facilities, etc.) are suppressed only when there are fewer than five year-round housing units in the geographic area being tabulated regardless of the number of occupied housing units or the number of persons.

Suppression of Family, Household, or Occupied Housing Characteristics. Characteristics of families, households, or occupied housing units are shown if there are at least five occupied housing units within the geographic area tabulated.

Suppression of Owner or Rental Characteristics. Distributions of data for owners or renters are shown only when the number of owners is at least five and the number of renters is also at least five.

Suppression of Tables Cross-Classified by Race or Spanish Origin. Population and housing characteristics cross-classified by race or Spanish origin are subject to an additional level of scrutiny. On this level the 15 person or five household criteria stated above are also applied to each race or Spanish origin category.

Individual cells of data for specific race or Spanish origin groups are not suppressed when there are 15 or more persons of that group in a geographic area.

The population and housing suppression criteria are applied independently of one another. For example, if there are 16 Spanish origin persons but only four households with Spanish origin householders, the person characteristics will be shown but the family, household, and housing characteristics will be suppressed.

#### Complementary Suppression

In some cases complementary suppression is applied to prevent the derivation of suppressed data by subtraction. For instance, when a table shows the number of persons in unit for all households and also for renters, there must be at least five owners and five renters for the renter data to be shown; otherwise the characteristics of the owners could be derived by subtracting renter data from data for all households.

#### Examples of Suppression--A Record

The following example shows two A record tables from the STF 2 Data Dictionary. The first Table (A13) is never suppressed since it is a basic count. The second table (A22) will only be suppressed if there are fewer than 5 year-round housing units in the geographic area tabulated.

#### Example:

Table A13. SPANISH ORIGIN (2) BY RACE (5)

This table has no suppression

Universe: Persons

This table has no suppression because a count of persons by race is never suppressed.

Spanish origin:  
White  
Black  
American Indian, Eskimo, and Aleut  
Asian and Pacific Islander  
Other

Not of Spanish origin:  
White  
Black  
American Indian, Eskimo, and Aleut  
Asian and Pacific Islander  
Other

Table A22. ROOMS (6)

SUPFLG02 applies to all cells

Universe: Condominium Housing Units

This table will be suppressed when there are 1-4 year-round housing units in the area.

1 room  
2 rooms  
3 rooms  
4 rooms  
5 rooms  
6 or more rooms

Examples of Suppression--B Record

The following examples show two B record tables from the STF 2 Data Dictionary. The first table (B8) will be suppressed if there are fewer than 15 persons of a particular race or Spanish origin category in the geographic area being summarized. The second table (B28) will have one portion of the table suppressed if there are fewer than five occupied housing units of a particular race or Spanish origin category in the geographic area, and the other portion suppressed if there are fewer than 5 owner and/or renter housing units of the same race or Spanish origin category in the geographic area.

Example:

Table B8. SEX (2) BY AGE (103)

SUPFLGB1 applies to all cells

Universe: Persons

This table will be suppressed when

Total:  
Under 1 year

there are 1-14  
persons in the  
area.

1 year  
2 year  
3 years  
"  
"  
to  
"  
"  
99 years  
100 to 104 years  
105 to 109 years  
110 years and over

Female:  
Under 1 year  
1 year  
2 years  
3 years  
"  
"  
to  
"  
"  
99 years  
100 to 104 years  
105 to 109 years  
110 years and over

Table B28. TENURE (2) BY UNITS AT ADDRESS (6)

SUPFLGB3 applies to cells 1-6

SUPFLGB4 applies to cells 7-12

Universe: Occupied Housing Units

The portion of this  
table indicating total  
will be suppressed  
when there are 1-4  
occupied housing units  
in the area.

Total:  
1  
2  
3 and 4  
5 to 9  
10 or more  
Mobile home or trailer

The portion of this  
table indicating renter  
occupied will be sup-

Renter occupied:  
1  
2



Pressed if there are 1-4  
owner and/or renter  
occupied housing units  
in the area.

3 and 4  
5 to 9  
10 or more  
Mobile home or trailer

How Suppression Affects the B Record in STF 2--Total suppression of the B record for a specified race or Spanish origin category occurs if the population in that particular category is less than 15 and the number of occupied housing units is less than 5.

B records may be partially suppressed because population and housing suppression criteria are applied independently of each other. For example, in the Asian and Pacific Islander category, if there are 16 persons but only 4 housing units for an area, the person characteristics will be shown but the family, household, and housing characteristics will be suppressed.

#### Programming with Suppression

Suppressed data cells contain zeroes. To distinguish between zeroes as suppression and zeroes as valid data, occurrences of suppression are identified by a series of flag fields in the geographic identification portion of each logical record. Programmers developing software should include procedures to check these fields for the presence of suppression and, if necessary, to flag the output of any cumulation which includes one or more suppressed fields.

In reviewing the data dictionary, the programmer can determine which suppression flags indicate suppression for particular tables by checking either the table description or the flag description. An example of each follows.

Example: The boxed illustration below is the table description as it appears in the data dictionary. The next portion illustrates the suppression flag to which the table description refers.

TABLE A32 CONTRACT RENT (26)

SUPFLG06 applies to all cells

SUPFLG06

Renter Occupied Housing Unit  
Suppression Flag

A 1 in this field indicates suppression because there are fewer than five renter housing units in the geographic area being summarized or complementary suppression is applied. It will affect the

following tables:

A29  
A32  
A33  
0 No suppression  
1 Suppression

Figure 10 below, lists each suppression flag, its location within the record, and the tables or cells within tables which are affected when suppression is applied. The suppression flag field which applies to each table or portion of a table is also identified in the table description in the data dictionary. The flags are located in the geographic identification section of each record in positions 205-210.

Figure 10. Suppression Flags

<u>Name</u>	<u>Begin</u>	<u>Table</u>
<u>Record Type A</u>		
SUPFLG01	205	A9, A10, A18, A21, A56 (cells 1-26), A60 (cells 1-18), A61 (cells 1-4), A62 (cells 1-10)
SUPFLG02	206	A22, A34, A35 (cells 1-9), A36 (cell 1), A37 (cell 1), A38 (cell 1), A39 (cells 1-4)
SUPFLG03	207	A2, A6 (cells 3-4), A14, A15 (cells 1-77), A16 (cells 1-14), A17 (cells 1-10), A19 (cells 1-42), A20, A23-25, A26 (cells 1-7), A35 (cells 19-45), A36 (cell 2), A36 (cells 5-8), A37 (cells 3-5), A38 (cells 3-5), A39 (cells 9-20)
SUPFLG04	208	A15 (cells 78-154), A16 (cells 15-28), A17 (cells 11-20), A19 (cells 43-84), A26 (cells 8-14), A35 (cells 10-18), A36 (cells 3-4), A37 (cell 2), A38 (cell 2), A39 (cells 5-8)
SUPFLG05	209	A27, A28, A30, A31
SUPFLG06	210	A29, A32, A33
<u>Record Type B</u>		
SUPFLGB1	205	B4-B13, B16, B17, B21

SUPFLGB2	206	-
SUPFLGB3	207	B2, B14, B15, B18, B19 (cell 1), B20 (cells 1-9), B22 (cell 1), B23 (cells 1-9), B24 (cell 1), B25 (cell 1), B26 (cells 1-4), B27 (cells 1-14), B28 (cells 1-6), B29 (cells 1-2), B30 (cells 1-2), B31 (cells 1-4), B32 (cells 1-4), B33 (cells 1-3), B34 (cells 1-3), B35 (cell 1), B36 (cell 1)
SUPFLGB4	208	B20 (cells 10-18), B22 (cells 2-3), B23 (cells 10-18), B24 B24 (cells 2-3), B25 (cell 2), B26 (cells 5-8), B27 (cells 15-28), B28 (cells 7-12), B29 (cells 3-4), B30 (cells 3-4), B31 (cells 5-8), B32 (cells 5-8), B33 (cells 4-6), B34 (cells 4-6), B35 (cell 2), B36 (cell 2)
SUPFLGB5	209	B19 (cell 4), B37-B39
SUPFLGB6	210	B19 (cells 2, 3, 5, 6), B40-B42

---

Evaluating the Effect of Suppression

In most cases, suppressed data values are small (fewer than 5 or 15) except where a large population is affected by complementary suppression. Therefore, in certain noncritical applications, users may simplify programming operations by ignoring suppression and treating suppressed cells as zero cells. However, when geographic entities are being summed to higher levels or new geographic areas are being created, suppression will usually result in a downward bias in the totals.

# APPENDIX B

## 1980, part 2

1980 Census of Housing and  
Population

User's Guide

PART A. TEXT

pp. 103-106

prepared by

the U.S. Census Bureau, 1982

**SUPPRESSION**

In order to maintain the confidentiality promised respondents and required by law, the Census Bureau withholds or "suppresses" tabulations of characteristics of very small groups of people or housing units.

In printed and microfiche reports, each suppressed data item is replaced by three dots (...), as illustrated in figure 6-11. On summary tape files, special flags denote suppressed data.

The suppression of certain data may inconvenience data users, especially when they are aggregating data for groups of blocks or tracts. The incon-

venience can be lessened if one understands the rules the Census Bureau followed in its disclosure analysis.

**Basic Principles Governing Suppression**

The Bureau never suppresses certain basic counts, even if an area has a count of only one. These basic counts are as follows:

- Total population
- Total housing units
- Year-round housing units
- Occupied units
- Vacant year-round housing units
- Counts of persons and households in each race and Spanish-origin category

All other data may be suppressed under certain conditions (discussed in detail below), primarily where the size of the population being characterized

is less than a specified threshold. The suppression criteria differ for population data and household data. Also, the thresholds are higher for sample estimates than for complete counts. The application of these thresholds results in what is known as "primary suppression." In addition, the Bureau applies "complementary suppression" to avoid the possibility of disclosure by subtraction.

**Suppression of Person Characteristics Derived from the Complete Count.** Complete counts of person characteristics other than race or Spanish origin (e.g., age or relationship) are shown only if there are 15 or more persons in the geographic area. For example, in data for a block with a population of 1 to 14 persons, population characteristics such as age and relationship are suppressed; tabulations show only counts for total population and the numbers of persons in specific race or Spanish-origin groups.

FIGURE 6-10 Allocation/Substitution Table Outlines from PC86-1-B

Table B-1. Characteristics of Persons Before and After Substitution and Allocation: 1980—Con.						
(For meaning of symbols, see introduction. For definitions of terms, see appendices A and B)						
The State	Persons			Percent		
	After substitution and allocation	After substitution	Before substitution and allocation	After substitution and allocation	After substitution	Before substitution and allocation
<b>HOUSEHOLD RELATIONSHIP</b>						
Total persons						
by household						
Married						
Spouse						
Child						
Elderly or other						
Parent						
Other relative						
Nonrelative						
by group quarters						
Inmate of institution						
Other						
<b>SEX</b>						

Table B-4. Percent of Substitution and Allocation: 1980—Con.										
(For meaning of symbols, see introduction. For definitions of terms, see appendices A and B)										
The State Urban and Rural and Size of Place Inside and Outside SMSA's SCSA's SMSA's Urbanized Areas Places of 1,000 or More Counties	Total persons (number)	Persons substituted for—		Persons with allocated—						
		Mechanical failure	Nonreview	Persons with one or more items allocated	Relationship	Sex	Age	Race	Origin	Married status— Persons 15 years and over
The State										
<b>URBAN AND RURAL AND SIZE OF PLACE</b>										
Urban										
Inside urbanized areas										
Central cities										
Urban fringe										
Outside urbanized areas										
Places of 10,000 or more										
Places of 2,500 to 10,000										
Rural										
Places of 1,000 to 2,500										
Other rural										
<b>INSIDE AND OUTSIDE SMSA's</b>										



The 15-person criterion applies only to the applicable "critical universe," in this case, total persons. These rules would not prevent the display of data showing, for example, that there are 2 persons 65 years old or over, as long as the area includes 15 or more persons in total.

**Suppression of Family, Household or Housing Unit Characteristics Derived from the Complete Count.** The threshold for family, household or housing unit data is 5, not 15. Characteristics of year-round housing units are shown if the area includes five or more year-round units. Characteristics of occupied housing units, households, or families are shown if the area includes five or more occupied units. Similar thresholds govern characteristics of owners and renters, except that the Bureau must also avoid complementary disclosure. For example, if an area includes 10 occupied housing units, 8 rented and 2 owner-occupied, any data provided for the total and for renters would be derivable for the 2 owners by subtraction. Therefore, most characteristics of owners and renters are shown only if the area includes at least five owners and five renters.

The suppression criteria for population and housing are applied independently of each other. For example, if an area includes 16 persons but only 4 housing units, the person characteristics will be shown but family, household and housing characteristics will be suppressed.

**Suppression of Complete-Count Tables Cross-Classified by Race or Spanish Origin.** Population and housing characteristics cross-classified by race or Spanish-origin are subject to an additional level of scrutiny. The 15-person or 5-household criteria stated above for complete-count data also apply to each race or Spanish-origin category. For example, a table of race by age for a geographic area that has 200 persons—124 White; 14 Black; 10 American Indian, Eskimo and Aleut; and 52 Asian and Pacific Islanders—shows actual age data for Whites and the Asian and Pacific Islander group, but not for the 2 groups with fewer than 15 persons.

On the other hand, if only one of the race categories in the foregoing example had more than zero but fewer than 15 persons, the Bureau would have employed complementary suppression to avoid the derivation of data about that one race by subtraction. Figure 6-12 illustrates the fact that a second race

would have been complementarily suppressed, generally the "other race" category (or "race, n.e.c." in sample data); but, since that group has no population in the example, the next smallest race group is targeted for complementary suppression. (Complementary suppression is not always obvious since most published tables omit the "other race" category, thereby requiring that it be derived by the subtraction of data for specified races from the total. If one of the specified races is suppressed, characteristics for the "other race" category can no longer be derived.)

**Suppression in Sample Data.** Thresholds applied to sample estimates are double those applied to complete counts, i.e., 30 persons/10 households instead of 15 persons/5 households. Otherwise, the rules are analogous. The size of the sample in the area (50 percent or 16 2/3 percent) does not affect the thresh-

olds. Note that it is a sample estimate that is tested relative to the threshold; for example, an area with 30 persons in the complete count but only 25 persons estimated in the sample would have its sample characteristics suppressed.

Suppression of sample data normally should be of less concern than complete-count suppression, since any sample number small enough to be suppressed would have been unreliable anyway.

**Illustrations of Suppression**

Users occasionally misunderstand census suppression rules since they expect suppression to be on a cell-by-cell basis (e.g., every number less than 15 suppressed) rather than a critical universe basis (e.g., category cells suppressed only if the total population is less than 15 persons).

FIGURE 6-12 Hypothetical Table Illustrating Suppression in Complete-Count Data

Race by Age	Data Before Suppression	Data As Made Public	
Total	200	200	
Under 5 years	10	10	
5 to 17 years	20	20	
18 to 64 years	140	140	
65 years and over	30	30	
White	124	124	
Under 5 years	7	7	
5 to 17 years	11	11	
18 to 64 years	90	90	
65 years and over	16	16	
Black	14	14	
Under 5 years	1	Suppressed	} Primary Suppression
5 to 17 years	1	Suppressed	
18 to 64 years	10	Suppressed	
65 years and over	2	Suppressed	
American Indian, Eskimo, and Aleut	62	62	
Under 5 years	2	Suppressed	} Complementary Suppression
5 to 17 years	8	Suppressed	
18 to 64 years	40	Suppressed	
65 years and over	12	Suppressed	
Asian and Pacific Islander	0	0	
Under 5 years	0	0	
5 to 17 years	0	0	
18 to 64 years	0	0	
65 years and over	0	0	
Other	0	0	
Under 5 years	0	0	
5 to 17 years	0	0	
18 to 64 years	0	0	
65 years and over	0	0	

Several aspects of these suppression criteria are illustrated by the following examples. The number of owners and the number of renters are critical universes for certain tabulations. For complete-count housing value data to be shown, an area must include at least five owners, and for rent data, at least five renters. For plumbing facilities data to be shown for renters, an area must include both five owners and five renters, since owner data would be derivable by subtracting renter data on plumbing facilities from corresponding data for all occupied units. On the other hand, if a table shows only the number of owners and renters (no characteristics), the only requirement is that there be at least five occupied units. A table cross-tabulating plumbing facilities and persons per room is also subject only to the five occupied unit threshold, since plumbing facilities and persons per room are not involved in the definition of critical universes.

The user need not memorize these criteria, only understand the general principles. More explicit detail about which criteria apply to which data cells is provided in summary tape technical documentation.

Certain reports show data for race or Spanish-origin groups only if the race or origin group in the given area meets a certain threshold (usually 400 or 1,000 persons). The purpose is not to avoid disclosure, but merely to reduce publication costs. Complete-count data are available on STF 2 for race or Spanish-origin groups with 15 or more persons and 5 or more households, and sample estimates are available on STF 4 for similar groups with 30 or more persons and 10 or more households—each in at least as much detail as is available for larger groups in print.

### Programming with Suppression

**Suppression Indicators.** Suppressed data cells on summary tape files contain zeroes. To distinguish between zeroes as suppression and zeroes as valid data, occurrences of suppression are identified by a series of flag fields in the geographic identification portion of each data record. Programmers developing software should include procedures to check these fields for the presence of suppression and, if necessary, to flag the output of any cumulation which includes one or more suppressed fields.

Technical documentation for each STF defines the relationship between

data tables and the suppression flags in two ways. First, the description of each flag, in the identification section of each record, lists each table or part which is governed by that suppression flag. Second, each table description indicates which suppression flag applies.

**Consequences of Ignoring Suppression.** In most cases, suppressed data values are small (less than 30 in any case). A sizable percentage of individual suppressed data cells were actual zeroes before suppression, although a large population may be affected by complementary suppression. Therefore, in certain applications that are not critical, users may simplify programming operations by ignoring suppression and treating suppressed cells as zero cells.

However, if the user is adding up blocks or enumeration districts to derive tables for specially defined areas, ignoring suppression will result in a downward bias in the totals. A user can gauge the impact of the downward bias if the universe of the tabulation is one that is never suppressed, as the following example illustrates. An age distribution for all persons may be suppressed, but the total number of persons is never suppressed. Therefore, if an age distribution is cumulated for a user-defined group of blocks, the total population should also be cumulated. If the sum of persons in all age categories for the group of blocks is 425 and the total population is 460, one can conclude that there were 35 people in blocks where the age distribution was suppressed.



# APPENDIX B

## 1980, part 3

1980 Census of Population  
General Population  
Characteristics  
UNITED STATES SUMMARY  
Introduction, pp. IV-V  
prepared by  
the U.S. Census Bureau, 1983

## SYMBOLS AND GEOGRAPHIC ABBREVIATIONS

The following symbols and geographic abbreviations are used in the tables:

- A dash "-" represents zero or a percent which rounds to less than 0.1.
- Three dots "... " mean not applicable, or that the data are being withheld to avoid disclosure of information for individuals or housing units. (For further information on disclosure, see the section below on "Suppression of Data for Confidentiality.")
- CDP is census designated place.
- SCSA is standard consolidated statistical area.
- SMSA is standard metropolitan statistical area.

## SUPPRESSION OF DATA FOR CONFIDENTIALITY

To maintain the confidentiality promised respondents and required by law, the Census Bureau takes precautions that its published data do not disclose information about specific individuals and housing units. To accomplish this, the Bureau suppresses data for characteristics which are based on a small number of persons and/or housing units in the geographic area. Under certain conditions, both primary and complementary suppression, as defined below, may take place.

The general rules of primary suppression of sample data are as follows: estimates of total population by race and Spanish origin are never suppressed; other characteristics for persons are shown only if there are 30 or more persons in the geographic area; estimates of total housing units, vacant housing units, year-round housing units, and occupied housing units are never suppressed; characteristics of year-round housing units which are not classified by occupancy status are shown only when there are 10 or more year-round housing units in the geographic area; characteristics of families, households, or occupied housing units are shown only if there are at least 10 occupied housing units within the geographic area; and distributions of data for owners or renters are shown only where the number of owners is at least 10 and the number of renters is also at least 10. These primary suppression criteria are applied independently of one another. The comparable figures for complete count (100 percent) data are 15 or more persons and 5 or more housing units of the specified type.

Population and occupied housing unit characteristics cross-classified by race or Spanish origin (of the householder in the case of occupied housing units) are subject to an additional level of examination. This requires the 30 person or 10 housing unit criterion stated above be applied individually to each race or Spanish origin category.

Finally, complementary suppression is applied to prevent the derivation of primary suppressed data by subtraction. For example, housing unit data shown by tenure may require complementary suppression when the number of owner-occupied or renter-occupied housing units is less than 10.

# APPENDIX C

## 1990

1990 Census of Population  
General Population  
Characteristics  
Delaware  
Appendix C, page 1  
prepared by  
the U.S. Census Bureau, 1992

Note that the same documentation appeared in  
all 1990 Summary Files.

## APPENDIX C. Accuracy of the Data

---

### CONTENTS

Confidentiality of the Data .....	C-1
Editing of Unacceptable Data .....	C-1
Sources of Error .....	C-1

### CONFIDENTIALITY OF THE DATA

To maintain confidentiality required by law (Title 13, United States Code), the Bureau of the Census applies a confidentiality edit to assure published data do not disclose information about specific individuals, households, and housing units. The result is that a small amount of uncertainty is introduced into some of the census characteristics to prevent identification of specific individuals, households, or housing units. The edit is controlled so that the counts of total persons, totals by race and American Indian tribe, Hispanic origin, and age 18 years and over are *not affected* by the confidentiality edit and are published as collected. In addition, total counts for housing units by tenure are not affected by this edit.

The confidentiality edit is conducted by selecting a sample of census households from the 100-percent data internal census files and interchanging its data with other households that have identical characteristics on a set of selected key variables but are in different geographic locations within the same State. To provide more protection for "small areas," a higher sampling rate was used for these areas. The net result of this procedure is that the data user's ability to obtain census data, particularly for small areas and subpopulation groups, has been significantly enhanced.

---

# APPENDIX D

## 2000

Census 2000 Summary File 3  
Technical Documentation  
Chapter 8, pp. 8-4 and 8-5  
prepared by  
the U.S. Census Bureau, 2002

Note that the same documentation appeared in  
all 2000 Summary Files.

#### CONFIDENTIALITY OF THE DATA

The Census Bureau has modified or suppressed some data in this data release to protect confidentiality. Title 13 United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual can be identified. The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases. A checklist approach is used to ensure that all potential risks to the confidentiality of the data are considered and addressed.

**Title 13, United States Code.** Title 13 of the United States Code authorizes the Census Bureau to conduct censuses and surveys. Section 9 of the same Title requires that any information collected from the public under the authority of Title 13 be maintained as confidential. Section 214 of Title 13 and Sections 3559 and 3571 of Title 18 of the United States Code provide for the imposition of penalties of up to 5 years in prison and up to \$250,000 in fines for wrongful disclosure of confidential census information.

8-4

Accuracy of the Data

U.S. Census Bureau, Census 2000

---

**Disclosure limitation.** Disclosure limitation is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify an individual who provided information under a pledge of confidentiality. Using disclosure limitation procedures, the Census Bureau modifies or removes the characteristics that put confidential information at risk for disclosure. Although it may appear that a table shows information about a specific individual, the Census Bureau has taken steps to disguise the original data while making sure the results are still useful. The techniques used by the Census Bureau to protect confidentiality in tabulations vary, depending on the type of data.

**Data swapping.** Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percentage of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases when creating a table. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics (such as the same number of adults and the same number of children). Because the swap often occurs within a neighboring area, there is no effect on the marginal totals for the area or for totals that include data from multiple areas. Because of data swapping, users should not assume that tables with cells having a value of 1 or 2 reveal information about specific individuals. Data swapping procedures were first used in the 1990 census and were also used for Census 2000.

# APPENDIX E

## 2010

2010 Census Advance  
Group Quarters Summary File--  
Technical Documentation  
Chapter 5, page 6  
prepared by the  
U.S. Census Bureau, 2011.

Note that the same documentation appeared in  
all 2010 Summary Files.

## **CONFIDENTIALITY OF THE DATA**

The Census Bureau has modified some data in this data release to protect confidentiality. Title 13 U.S. Code, Section 9, prohibits the Census Bureau from publishing results in which an individual's data can be identified.

The Census Bureau's internal Disclosure Review Board monitors the disclosure review process and sets the confidentiality rules for all data releases. A checklist approach is used to ensure that all potential risks are considered and addressed. A list of possible concerns is created and the Disclosure Review Board makes sure that the appropriate steps are taken to assure the confidentiality of the data.

### **Title 13 U.S. Code**

Title 13 of the U.S. Code authorizes the Census Bureau to conduct surveys and censuses and mandates that any information obtained from private individuals and establishments remains confidential. Section 9 of Title 13 prohibits the Census Bureau from releasing "any publication whereby the data furnished by any particular establishment or individual under this title can be identified." Section 214 of Title 13, as modified by the Federal Sentencing Reform Act, imposes a fine of not more than \$250,000 and/or imprisonment of not more than 5 years for publication or communication in violation of Section 9.

### **Disclosure Avoidance**

Disclosure avoidance is the process of disguising data to protect confidentiality. A disclosure of data occurs when someone can use published statistical information to identify an individual who provided information under a pledge of confidentiality. Using disclosure avoidance, the Census Bureau modifies or removes all of the characteristics that put confidential information at risk for disclosure. Although it may appear that a table shows information about a specific individual, the Census Bureau has taken steps (such as data swapping) to disguise the original data while making sure the results are useful.

### **Data Swapping**

Data swapping is a method of disclosure avoidance designed to protect confidentiality in tables of frequency data (the number or percentage of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas (geographic areas with a small population) that have similar characteristics (same number of adults, same number of children, etc.). Because the swap often occurs within a geographic area with a small population, there is no effect on the marginal totals for the geographic area with a small population or for totals that include data from multiple geographic areas with small populations. Because of data swapping, users should not assume that tables with cells having a value of one or two reveal information about specific individuals.





SCIENTIFIC AMERICAN

Subscribe

Read Our Latest Issue [Read Now](#) X

HEALTH

# Privacy by the Numbers: A New Approach to Safeguarding Data

A mathematical technique called “differential privacy” gives researchers access to vast repositories of personal data while meeting a high standard for privacy protection

By Erica Klarreich, Quanta Magazine on December 31, 2012

From ~~Simons Science News~~ (find original story here)

In 1997, when Massachusetts began making health records of state employees available to medical researchers, the government removed patients’ names, addresses, and Social Security numbers. William Weld, then the governor, assured the public that identifying individual patients in the records would be impossible.

Within days, an envelope from a graduate student at the Massachusetts Institute of Technology arrived at Weld’s office. It contained the governor’s health records.





ADVERTISEMENT

Although the state had removed all obvious identifiers, it had left each patient's date of birth, sex and ZIP code. By cross-referencing this information with voter-registration records, Latanya Sweeney was able to pinpoint Weld's records.

Sweeney's work, along with other notable privacy breaches over the past 15 years, has raised questions about the security of supposedly anonymous information.

"We've learned that human intuition about what is private is not especially good," said Frank McSherry of Microsoft Research Silicon Valley in Mountain View, Calif.

"Computers are getting more and more sophisticated at pulling individual data out of things that a naive person might think are harmless."

As awareness of these privacy concerns has grown, many organizations have clamped down on their sensitive data, uncertain about what, if anything, they can release without jeopardizing the privacy of individuals. But this attention to privacy has come at a price, cutting researchers off from vast repositories of potentially invaluable data.

Medical records, like those released by Massachusetts, could help reveal which genes increase the risk of developing diseases like Alzheimer's, how to reduce medical errors in hospitals or what treatments are most effective against breast cancer. Government-held information from Census Bureau surveys and tax returns could help economists devise policies that best promote income equality or economic growth. And data from social

media websites like Facebook and Twitter could offer sociologists an unprecedented look at how ordinary people go about their lives.

ADVERTISEMENT

The question is: How do we get at these data without revealing private information? A body of work a decade in the making is now starting to offer a genuine solution.

“Differential privacy,” as the approach is called, allows for the release of data while meeting a high standard for privacy protection. A differentially private data release algorithm allows researchers to ask practically any question about a database of sensitive information and provides answers that have been “blurred” so that they reveal virtually nothing about any individual’s data — not even whether the individual was in the database in the first place.

“The idea is that if you allow your data to be used, you incur no additional risk,” said Cynthia Dwork of Microsoft Research Silicon Valley. Dwork introduced the concept of differential privacy in 2005, along with McSherry, Kobbi Nissim of Israel’s Ben-Gurion University and Adam Smith of Pennsylvania State University.



Sign up for *Scientific American's* free newsletters.

Sign Up

Differential privacy preserves “plausible deniability,” as Avrim Blum of Carnegie Mellon University likes to put it. “If I want to pretend that my private information is different from what it really is, I can,” he said. “The output of a differentially private mechanism is going to be almost exactly the same whether it includes the real me or the pretend me, so I can plausibly deny anything I want.”

This privacy standard may seem so high as to be unattainable — and indeed, there is no differentially private algorithm that gives out *exactly* the same information regardless of whether it includes the real you or the pretend you. But if we allow



algorithms that give out *almost* exactly the same information in the two cases, then useful and efficient algorithms do exist. This “almost” is a precisely calibrated parameter, a measurable quantification of privacy. Individuals or social institutions could decide what value of this parameter represents an acceptable loss of privacy, and then differentially private algorithms could be chosen that guarantee that the privacy loss is less than the selected parameter.

#### ADVERTISEMENT

Privacy experts have developed a wide assortment of specialized differentially private algorithms to handle different kinds of data and questions about the data. Although much of this work is technical and difficult for nonexperts to penetrate, researchers are starting to build standardized computer languages that would allow nonexperts to release sensitive data in a differentially private way by writing a simple computer program.

One real-world application already uses differential privacy: a Census Bureau project called *OnTheMap*, which gives researchers access to agency data. Also, differential privacy researchers have fielded preliminary inquiries from Facebook and the federally funded iDASH center at the University of California, San Diego, whose mandate in large part is to find ways for researchers to share biomedical data without compromising privacy.

“Differential privacy is a promising and exciting technology,” said Aaron Roth, a computer scientist at the University of Pennsylvania.

### **Needle in a Haystack**

It might seem that a simpler solution to the privacy problem would be to release only “aggregate” information — statements about large groups of people. But even this approach is susceptible to breaches of privacy.

Suppose you wanted to ascertain whether this writer has diabetes and you knew I



logged to a health database. You could find this out simply by subtracting the answers

to two aggregate-level questions: “How many people in the database have diabetes?” and “How many people in the database not named Erica Klarreich have diabetes?”

ADVERTISEMENT

Clearly, these two questions, when combined, violate my privacy. But it’s not always easy to spot which combinations of questions would constitute privacy breaches. Spotting such combinations is, in its full generality, what computer scientists call an “NP-hard” problem, which means that there is probably no efficient computer algorithm that could catch all such attacks.

And when the attacker has access to outside information about individuals in the database, extracting private information from aggregate statistics becomes even easier.

In 2008, a research team demonstrated the dangers of releasing aggregate information from genome-wide association studies, one of the primary research vehicles for uncovering links between diseases and particular genes. These studies typically involve sequencing the genomes of a test group of 100 to 1,000 patients who have the same disease and then calculating the average frequency in the group of something on the order of 100,000 different mutations. If a mutation appears in the group far more frequently than in the general population, that mutation is flagged as a possible cause or contributor to the disease.

The research team, led by Nils Homer, then a graduate student at the University of California at Los Angeles, showed that in many cases, if you know a person’s genome, you can figure out beyond a reasonable doubt whether that person has participated in a particular genome-wide test group. After Homer’s paper appeared, the National Institutes of Health reversed a policy, instituted earlier that year, that had required aggregate data from all NIH-funded genome-wide association studies to be posted publicly.

Perhaps even more surprisingly, researchers showed in 2011 that it is possible to glean personal information about purchases from Amazon.com’s product recommendation system, which makes aggregate-level statements of the form, “Customers who bought



this item also bought A, B and C.” By observing how the recommendations changed over time and cross-referencing them with customers’ public reviews of purchased items, the researchers were able in several cases to infer that a particular customer had bought a particular item on a particular day — even before the customer had posted a review of the item.

#### ADVERTISEMENT

In all these cases, the privacy measures that had been taken seemed adequate, until they were breached. But even as the list of privacy failures ballooned, a different approach to data release was in the making, one that came with an a priori privacy guarantee. To achieve this goal, researchers had gone back to basics: Just what does it mean, they asked, to protect privacy?

### **Two-World Privacy**

If researchers study a health database and discover a link between smoking and some form of cancer, differential privacy will not protect a public smoker from being labeled with elevated cancer risk. But if a person’s smoking is a secret hidden in the database, differential privacy will protect that secret.

“Differential’ refers to the difference between two worlds — one in which you allow your sensitive data to be included in the database and one in which you don’t,” McSherry said. The two worlds cannot be made to work out exactly the same, but they can be made close enough that they are effectively indistinguishable. That, he said, is the goal of differential privacy.

Differential privacy focuses on information-releasing algorithms, which take in questions about a database and spit out answers — not exact answers, but answers that have been randomly altered in a prescribed way. When the same question is asked of a pair of databases (*A* and *B*) that differ only with regard to a single individual (Person *X*), the algorithm should spit out essentially the same answers.



precisely, given any answer that the algorithm could conceivably spit out, the probability of getting that answer should be almost exactly the same for both databases;

that is, the ratio of these two probabilities should be bounded by some number  $R$  close to 1. The closer  $R$  is to 1, the more difficult it will be for an attacker to figure out whether he is getting information about database  $A$  or database  $B$  and the better protected Person  $X$  will be. After all, if the attacker can't even figure out whether the information he is getting includes Person  $X$ 's data, he certainly can't figure out what Person  $X$ 's data is.

(Differential privacy researchers usually prefer to speak in terms of the logarithm of  $R$ , which they denote  $\epsilon$ . This parameter puts a number on how much privacy leaks out when the algorithm is carried out: The closer  $\epsilon$  is to 0, the better the algorithm is at protecting privacy.)

To get a sense of how differentially private algorithms can be constructed, let's look at one of the simplest such algorithms. It focuses on a scenario in which a questioner is limited to "counting queries"; for example: "How many people in the database have property  $P$ ?"

Suppose the true answer to one such question is 157. The differentially private algorithm will "add noise" to the true answer; that is, before returning an answer, it will add or subtract from 157 some number, chosen randomly according to a predetermined set of probabilities. Thus, it might return 157, but it also might return 153, 159 or even 292. The person who asked the question knows which probability distribution the algorithm is using, so she has a rough idea of how much the true answer has likely been distorted (otherwise the answer the algorithm spat out would be completely useless to her). However, she doesn't know which random number the algorithm actually added.

The particular probability distribution being used must be chosen with care. To see what kind of distribution will ensure differential privacy, imagine that a prying questioner is trying to find out whether I am in a database. He asks, "How many people named Erica Klarreich are in the database?" Let's say he gets an answer of 100. Because Erica Klarreich is such a rare name, the questioner knows that the true answer is almost certainly either 0 or 1, leaving two possibilities:



The answer is 0 and the algorithm added 100 in noise; or

(b) The answer is 1 and the algorithm added 99 in noise.

To preserve my privacy, the probability of picking 99 or 100 must be almost exactly the same; then the questioner will be unable to distinguish meaningfully between the two possibilities. More precisely, the ratio of these two probabilities should be at most the preselected privacy parameter  $R$ . And this should be the case with regard to not only 99 and 100 but also any pair of consecutive numbers; that way, no matter what noise value is added, the questioner won't be able to figure out the true answer.

A probability distribution that achieves this goal is the Laplace distribution, which comes to a sharp peak at 0 and gradually tapers off on each side. A Laplace distribution has exactly the property we need: There is some number  $R$  (called the width of the distribution) such that for any two consecutive numbers, the ratio of their probabilities is  $R$ .

There is one Laplace distribution for each possible width; thus, we can tinker with the width to get the Laplace distribution that gives us the exact degree of privacy we want. If we need a high level of privacy, the corresponding distribution will be comparatively wide and flat; numbers distant from 0 will be almost as probable as numbers close to 0, ensuring that the data are blurred by enough noise to protect privacy.

Inevitably, tension exists between privacy and utility. The more privacy you want, the more Laplace noise you have to add and the less useful the answer is to researchers studying the database. With a Laplace distribution, the expected amount of added noise is the reciprocal of  $R$ ; so, for example, if you have chosen a privacy parameter of 0.01, you can expect the algorithm's answer to be blurred by about 100 in noise.

The larger the dataset, the less a given amount of blurring will affect utility: Adding 100 in noise will blur an answer in the hundreds much more than an answer in the millions. For datasets on the scale of the Internet — that is, hundreds of millions of entries — the algorithm already provides good enough accuracy for many practical settings, Dwork said.





And the Laplace noise algorithm is only the first word when it comes to differential privacy. Researchers have come up with a slew of more sophisticated differentially private algorithms, many of which have a better utility-privacy trade-off than the Laplace noise algorithm in certain situations.

“People keep finding better ways of doing things, and there is still plenty more room for improvement,” Dwork said. When it comes to more moderate-sized datasets than the Internet, she said, “there are starting to be algorithms out there for many tasks.”

With a differentially private algorithm, there’s no need to analyze a question carefully to determine whether it seeks to invade an individual’s privacy; that protection is automatically built into the algorithm’s functioning. Because prying questions usually boil down to small numbers related to specific people and non-prying questions examine aggregate-level behavior of large groups, the same amount of added noise that renders answers about individuals meaningless will have only a minor effect on answers to many legitimate research questions.

With differential privacy, the kinds of issues that plagued other data releases — such as attackers cross-referencing data with outside information — disappear. The approach’s mathematical privacy guarantees do not depend on the attacker having limited outside information or resources.

“Differential privacy assumes that the adversary is all-powerful,” McSherry said. “Even if attackers were to come back 100 years later, with 100 years’ worth of thought and information and computer technology, they still would not be able to figure out whether you are in the database. Differential privacy is future-proofed.”

### **A Fundamental Primitive**

So far, we have focused on a situation in which someone asks a single counting query about a single database. But the real world is considerably more complex.

Researchers typically want to ask many questions about a database. And over your lifetime, snippets of your personal information will probably find their way into many different databases, each of which may be releasing data without consulting the others.



Differential privacy provides a precise and simple way to quantify the cumulative privacy hit you sustain if researchers ask multiple questions about the databases to which you belong. If you have sensitive data in two datasets, for example, and the curators of the two datasets release those data using algorithms whose privacy parameters are 1 and 2, respectively, then the total amount of your privacy that has leaked out is at most  $1 + 2$ . The same additive relationship holds if a curator allows multiple questions about a single database. If researchers ask  $m$  questions about a database and each question gets answered with privacy parameter  $\epsilon$ , the total amount of privacy lost is at most  $m\epsilon$ .

So, in theory, the curator of a dataset could allow researchers to ask as many counting queries as he wishes, as long as he adds enough Laplace noise to each answer to ensure that the total amount of privacy that leaks out is less than his preselected privacy “budget.”

And although we have limited our attention to counting queries, it turns out that this restriction is not very significant. Many of the other question types that researchers like to ask can be recast in terms of counting queries. If you wanted to generate a list of the top 100 baby names for 2012, for example, you could ask a series of questions of the form, “How many babies were given names that start with A?” (or Aa, Ab or Ac), and work your way through the possibilities.

“One of the early results in machine learning is that almost everything that is possible in principle to learn can be learned through counting queries,” Roth said. “Counting queries are not isolated toy problems, but a fundamental primitive” — that is, a building block from which many more complex algorithms can be built.

But there’s a catch. The more questions we want to allow, the less privacy each question is allowed to use up from the privacy budget and the more noise has to be added to each answer. Consider the baby names question. If we decide on a total privacy budget of 0.01 and there are 10,000 names to ask about, each question’s individual privacy budget is only  $0.01/10,000$ , or  $0.000001$ . The expected amount of noise added to each answer will be  $10,000/0.000001$ , or 1,000,000 — an amount that will swamp the true answers.




In other words, the naive approach of adding Laplace noise to each question independently is limited in terms of the number of questions to which it can provide useful answers. To deal with this, computer scientists have developed an arsenal of more powerful primitives — algorithmic building blocks which, by taking into account the particular structure of a database and problem type, can answer more questions with more accuracy than the naive approach can.

For example, In 2005, Smith noticed that the baby names problem has a special structure: removing one person’s personal information from the database changes the answer for only one of the 10,000 names in the database. Because of this attribute, we can get away with adding only  $1/10,000$  in Laplace noise to each name answer, instead of  $10,000/$ , and the outcome will stay within our privacy budget. This algorithm is a primitive that can be applied to any “histogram” query — that is, one asking how many people fall into each of several mutually exclusive categories, such as first names.

When Smith told Dwork about this insight in the early days of differential privacy research, “something inside me went, ‘Wow!’” Dwork said. “I realized that we could exploit the structure of a query or computation to get much greater accuracy than I had realized.”

Since that time, computer scientists have developed a large library of such primitives. And because the additive rule explains what happens to the privacy parameter when algorithms are combined, computer scientists can assemble these building blocks into complex structures while keeping tabs on just how much privacy the resulting algorithms use up.

“One of the achievements in this area has been to come up with algorithms that can handle a very large number of queries with a relatively small amount of noise,” said Moritz Hardt of IBM Research Almaden in San Jose, Calif.

To make differential privacy more accessible to nonexperts, several groups are working to create a differential privacy programming language that would abstract away all the underlying mathematics of the algorithmic primitives to a layer that the user doesn’t  
 to think about.

“If you’re the curator of a dataset, you don’t have to worry about what people are doing with your dataset as long as they are running queries written in this language,” said McSherry, who has created one preliminary such language, called PINQ. “The program serves as a proof that the query is OK.”

### **A Nonrenewable Resource**

Because the simple additive rule gives a precise upper limit on how much total privacy you lose when the various databases you belong to release information in a differentially private way, the additive rule turns privacy into a “fungible currency,” McSherry said.

For example, if you were to decide how much total lifetime privacy loss would be acceptable to you, you could then decide how you want to “spend” it — whether in exchange for money, perhaps, or to support a research project you admire. Each time you allowed your data to be used in a differentially private data release, you would know exactly how much of your privacy budget remained.

Likewise, the curator of a dataset of sensitive information could decide how to spend whatever amount of privacy she had decided to release — perhaps by inviting proposals for research projects that would describe not only what questions the researchers wanted to ask and why, but also how much privacy the project would use up. The curator could then decide which projects would make the most worthwhile use of the dataset’s predetermined privacy budget. Once this budget had been used up, the dataset could be closed to further study.

“Privacy is a nonrenewable resource,” McSherry said. “Once it gets consumed, it is gone.”

The question of which value represents an acceptable privacy loss is ultimately a problem for society, not for computer scientists — and each person may give a different answer. And although the prospect of putting a price on something as intangible as privacy may seem daunting, a relevant analog exists.

“There’s another resource that has the same property — the hours of your life,” McSherry said. “There are only so many of them, and once you use them, they’re gone.”



Yet because we have a currency and a market for labor, as a society we have figured out how to price people’s time. We could imagine the same thing happening for privacy.”

*Reprinted with permission from Simons Science News, an editorially independent division of SimonsFoundation.org whose mission is to enhance public understanding of science by covering research developments and trends in mathematics and the computational, physical and life sciences.*

---

**ABOUT THE AUTHOR(S)**

**Recent Articles by Erica Klarreich**

- The Mathematics of Cake Cutting
- Mathematicians Chase Moonshine’s Shadow
- A Fluid New Path in Grand Math Challenge

---

**Recent Articles by Quanta Magazine**

- To Invent a Quantum Internet
- A New "Law" Suggests Quantum Supremacy Could Happen This Year
- Icefish Study Adds Another Color to the Story of Blood

**READ THIS NEXT**

**SPONSORED**

How Oncologists use AI to Deliver Personalized Cancer Care

---

**EARTH**

Don't Worry about CO2, Worry about the Earth's 'Energy Balance'

By Leah Harvey and E&E News



**PHYSICS**

**Can Science Survive the Death of the Universe?**

John Horgan | Opinion

---

**PUBLIC HEALTH**

**The COVID Lab-Leak Hypothesis: What Scientists Do and Do Not Know**

Amy Maxmen, Smriti Mallapaty and Nature magazine

---

**EVOLUTION**

**Animal Kids Listen to Their Parents Even before Birth**

Karen Hopkin

---

**PUBLIC HEALTH**

**Labor Department Issues Emergency Rules to Protect Health Care Workers From COVID**

Christina Jewett and Kaiser Health News

---

**NEWSLETTER**

*Get smart. Sign up for our email newsletter.*

Sign Up

---

*Support Science Journalism*

Subscribe Now!



**FOLLOW US**



## SCIENTIFIC AMERICAN ARABIC

العربية

---

[Return & Refund Policy](#)

[FAQs](#)

[About](#)

[Contact Us](#)

[Press Room](#)

[Site Map](#)

[Advertise](#)

[Privacy Policy](#)

[SA Custom Media](#)

[California Consumer Privacy Statement](#)

[Terms of Use](#)

[Use of cookies/Do not sell my data](#)

[International Editions](#)

Scientific American is part of Springer Nature, which owns or has commercial relations with thousands of scientific publications (many of them can be found at [www.springernature.com/us](http://www.springernature.com/us)). Scientific American maintains a strict policy of editorial independence in reporting developments in science to our readers.

© 2021 SCIENTIFIC AMERICAN, A DIVISION OF SPRINGER NATURE AMERICA, INC.

ALL RIGHTS RESERVED.



# A Statistical Framework for Differential Privacy<sup>1</sup>

Larry Wasserman<sup>\*‡</sup> Shuheng Zhou<sup>†</sup>

<sup>\*</sup>Department of Statistics

<sup>‡</sup>Machine Learning Department

Carnegie Mellon University

Pittsburgh, PA 15213

<sup>†</sup>Seminar für Statistik

ETH Zürich, CH 8092

October 22, 2018

One goal of statistical privacy research is to construct a data release mechanism that protects individual privacy while preserving information content. An example is a *random mechanism* that takes an input database  $X$  and outputs a random database  $Z$  according to a distribution  $Q_n(\cdot|X)$ . *Differential privacy* is a particular privacy requirement developed by computer scientists in which  $Q_n(\cdot|X)$  is required to be insensitive to changes in one data point in  $X$ . This makes it difficult to infer from  $Z$  whether a given individual is in the original database  $X$ . We consider differential privacy from a statistical perspective. We consider several data release mechanisms that satisfy the differential privacy requirement. We show that it is useful to compare these schemes by computing the rate of convergence of distributions and densities constructed from the released data. We study a general privacy method, called the exponential mechanism, introduced by McSherry and Talwar (2007). We show that the accuracy of this method is intimately linked to the rate at which the probability that the empirical distribution concentrates in a small ball around the true distribution.

## 1 Introduction

One goal of data privacy research is to derive a mechanism that takes an input database  $X$  and releases a transformed database  $Z$  such that individual privacy is protected yet information content is preserved. This is known as disclosure limitation. In this paper we will consider various methods

---

<sup>1</sup> We thank Avrim Blum, Katrina Ligett, Steve Fienberg, Alessandro Rinaldo and Yuval Nardi for many helpful discussions. We thank Wenbo Li and Mikhail Lifshits for helpful pointers and discussions on small ball probabilities. We thank the Associate Editor and three referees for a plethora of comments that led to improvements in the paper. Research supported by NSF grant CCF-0625879, a Google research grant and a grant from Carnegie Mellon's Cylab. The second author is also partially supported by the Swiss National Science Foundation (SNF) Grant 20PA21-120050/1.



for producing a transformed database  $Z$  and we will study the accuracy of inferences from  $Z$  under various loss functions.

There are numerous approaches to this problem. The literature is vast and includes papers from computer science, statistics and other fields. The terminology also varies considerably. We will use the terms “disclosure limitation” and “privacy guarantee” interchangeably.

Disclosure limitation methods include clustering (Sweeney, 2002, Aggarwal et al., 2006),  $\ell$ -diversity (Machanavajjhala et al., 2006),  $t$ -closeness (Li et al., 2007), data swapping (Fienberg and McIntyre, 2004), matrix masking (Ting et al., 2008), cryptographic approaches (Pinkas, 2002, Feigenbaum et al., 2006), data perturbation (Evfimievski et al., 2004, Kim and Winkler, 2003, Warner, 1965, Fienberg et al., 1998) and distributed database methods (Fienberg et al., 2007, Sanil et al., 2004). Statistical references on disclosure risk and limitation include Duncan and Lambert (1986, 1989), Duncan and Pearson (1991), Reiter (2005). We refer to Reiter (2005) and Sanil et al. (2004) for further references.

One approach to defining a privacy guarantee that has received much attention in the computer science literature is known as *differential privacy* (Dwork et al., 2006, Dwork, 2006). There is a large body of work on this topic including, for example, Dinur and Nissim (2003), Dwork and Nissim (2004), Blum et al. (2005), Dwork et al. (2007), Nissim et al. (2007), Barak et al. (2007), McSherry and Talwar (2007), Blum et al. (2008), Kasiviswanathan et al. (2008). Blum et al. (2008) gives a machine learning approach to inference under differential privacy constraints and to some extent our results are inspired by that paper. Smith (2008) shows how to provide efficient point estimators while preserving differential privacy. He constructs estimators for parametric models with mean squared error  $(1 + o(1))/(nI(\theta))$  where  $I(\theta)$  is the Fisher information. Machanavajjhala et al. (2008) consider privacy for histograms by sampling from the posterior distribution of the cell probabilities. We discuss Machanavajjhala et al. (2008) further in Section 4. After submitting the first draft of this paper, new work has appeared on differential privacy that is also statistical in nature, namely, Ghosh et al. (2009), Dwork and Lei (2009), Dwork et al. (2009), Feldman et al. (2009).

The goals of this paper are to explain differential privacy in statistical language, to show how to compare different privacy mechanisms by computing the rate of convergence of distributions and densities based on the released data  $Z$ , and to study a general privacy method, called the exponen-

tial mechanism, due to McSherry and Talwar (2007). We show that the accuracy of this method is intimately linked to the rate at which the probability that the empirical distribution concentrates in a small ball around the true distribution. These so called “small ball probabilities” are well-studied in probability theory. To the best of our knowledge, this is the first time a connection has been made between differential privacy and small ball probabilities. We need to make two disclaimers. First, the goal of our paper is to investigate differential privacy. We will not attempt to review all approaches to privacy or to compare differential privacy with other approaches. Such an undertaking is beyond the scope of this paper. Second, we focus only on statistical properties here. We shall not concern ourselves in this paper with computational efficiency.

In Section 2 we define differential privacy and provide motivation for the definition. In Section 3 we discuss conditions that ensure that a privacy mechanism preserves information. In Section 4 we consider two histogram based methods. In Section 5 and 6, we examine another method known as the exponential mechanism. Section 7 contains a small simulation study and Section 8 contains concluding remarks. All technical proofs appear in Section 9.

## 1.1 Summary of Results

We consider several different data release mechanisms that satisfy differential privacy. We evaluate the utility of these mechanisms by evaluating the rate at which  $d(P, P_Z)$  goes to 0, where  $P$  is the distribution of the data  $X \in \mathcal{X}$ ,  $P_Z$  is the empirical distribution of the released data  $Z$ , and  $d$  is some distance between distributions. This gives an informative way to compare data release mechanisms. In more detail, we consider the Kolmogorov-Smirnov (KS) distance:  $\sup_{x \in \mathcal{X}} |F(x) - \widehat{F}_Z(x)|$ , where  $F, \widehat{F}_Z$  denote the cumulative distribution function (cdf) corresponding to  $P$  and the empirical distribution function corresponding to  $P_Z$ , respectively. We also consider the squared  $L_2$  distance:  $\int (p(x) - \widehat{p}_Z)^2$ , where  $\widehat{p}_Z$  is a density estimator based on  $Z$ . Our results are summarized in the following tables, where  $n$  denotes the sample size.

The first table concerns the case where the data are in  $\mathbb{R}^r$  and the density  $p$  of  $P$  is Lipschitz. Also reported are the minimax rates of convergence for density estimators in KS and in squared  $L_2$  distances. We see that the accuracy depends both on the data releasing mechanism and the distance

function  $d$ . The results are from Sections 4 and 5 of the paper. (The exponential mechanism under  $L_2$  distance is marked NA but is in the second table in case  $r = 1$ . We note that the rate for KS distance for perturbed histogram is  $\sqrt{\log n/n}$  for  $r = 1$ .)

Distance	Data Release mechanism			minimax rate
	smoothed histogram	perturbed histogram	exponential mechanism	
$L_2$	$n^{-2/(2r+3)}$	$n^{-2/(2+r)}$	NA	$n^{-2/(2+r)}$
Kolmogorov-Smirnov	$\sqrt{\log n} \times n^{-2/(6+r)}$	$\log n \times n^{-2/(2+r)}$	$n^{-1/3}$	$n^{-1/2}$

The next table summarizes the results for the case where the dimension of  $X$  is  $r = 1$  and the density  $p$  is assumed to be in a Sobolev space of order  $\gamma$ . We only consider the squared  $L_2$  distance between the true density  $p$  and the estimated density  $\hat{p}_Z$  in this case. The results are from Section 6 of the paper.

	exponential mechanism	perturbed orthogonal series estimator	minimax rate
$L_2$	$n^{-\gamma/(2\gamma+1)}$	$n^{-2\gamma/(2\gamma+1)}$	$n^{-2\gamma/(2\gamma+1)}$

Our results show that, in general, privacy schemes seem not to yield minimax rates. Two exceptions are perturbation methods evaluated under  $L_2$  loss which do yield minimax rates. An open question is whether the slower than minimax rates are intrinsic to the privacy methods. It is possible, for example, that our rates are not tight. This question could be answered by establishing lower bounds on these rates. We consider this an important topic for future research.

## 2 Differential Privacy

Let  $X_1, \dots, X_n$  be a random sample (independent and identically distributed) of size  $n$  from a distribution  $P$  where  $X_i \in \mathcal{X}$ . To be concrete, we shall assume that  $\mathcal{X} \equiv [0, 1]^r = [0, 1] \times [0, 1] \times \dots \times [0, 1]$  for some integer  $r \geq 1$ . Extensions to more general sample spaces are certainly possible but we focus on this sample space to avoid unnecessary technicalities. (In particular, it

is difficult to extend differential privacy to unbounded domains.) Let  $\mu$  denote Lebesgue measure and let  $p = dP/d\mu$  if the density exists. We call  $X = (X_1, \dots, X_n)$  a database. Note that  $X \in \mathcal{X}^n = [0, 1]^r \times \dots \times [0, 1]^r$ . We focus on mechanisms that take a database  $X$  as input and output a sanitized database  $Z = (Z_1, \dots, Z_k) \in \mathcal{X}^k$  for public release. In general,  $Z$  need not be the same size as  $X$ . For some schemes, we shall see that large  $k$  can lead to low privacy and high accuracy while while small  $k$  can lead to high privacy and low accuracy. We will let  $k \equiv k(n)$  change with  $n$ . Hence, any asymptotic statements involving  $n$  increasing will also allow  $k$  to change as well.

A *data release mechanism*  $Q_n(\cdot|X)$  is a conditional distribution for  $Z = (Z_1, \dots, Z_k)$  given  $X$ . Thus,  $Q_n(B|X = x)$  is the probability that the output database  $Z$  is in a set  $B \in \mathcal{B}$  given that the input database is  $x$ , where  $\mathcal{B}$  are the measurable subsets of  $\mathcal{X}^k$ . We call  $Z = (Z_1, \dots, Z_k)$  a *sanitized database*. Schematically:

$$\text{input database } X = (X_1, \dots, X_n) \xrightarrow[\text{sanitize}]{Q_n(Z|X)} \text{output database } Z = (Z_1, \dots, Z_k).$$

The marginal distribution of the output database  $Z$  induced by  $P$  and  $Q_n$  is  $M_n(B) = \int Q_n(B|X = x)dP^n(x)$  where  $P^n$  is the  $n$ -fold product measure of  $P$ .

**Example 2.1.** A simple example to help the reader have a concrete example in mind is adding noise. In this case,  $Z = (Z_1, \dots, Z_n)$  where  $Z_i = X_i + \epsilon_i$  and  $\epsilon_1, \dots, \epsilon_n$  are mean 0 independent observations drawn from some known distribution  $H$  with density  $h$ . Hence  $Q_n$  has density  $q_n(z_1, \dots, z_n|x_1, \dots, x_n) = \prod_{i=1}^n h(z_i - x_i)$ .

**Definition 2.2.** Given two databases  $X = (X_1, \dots, X_n)$  and  $Y = (Y_1, \dots, Y_n)$ , let  $\delta(X, Y)$  denote the Hamming distance between  $X$  and  $Y$ :  $\delta(X, Y) = \#\{i : X_i \neq Y_i\}$ .

A general data release mechanism is the *exponential mechanism* (McSherry and Talwar, 2007) which is defined as follows. Let  $\xi : \mathcal{X}^n \times \mathcal{X}^k \rightarrow [0, \infty)$  be any function. Each such  $\xi$  defines a different exponential mechanism. Let

$$\Delta \equiv \Delta_{n,k} = \sup_{\substack{x, y \in \mathcal{X}^n \\ \delta(x, y) = 1}} \sup_{z \in \mathcal{X}^k} |\xi(x, z) - \xi(y, z)|, \quad (1)$$

that is,  $\Delta_{n,k}$  is the maximum change to  $\xi$  caused by altering a single entry in  $x$ . Finally, let  $(Z_1, \dots, Z_k)$  be a random vector drawn from the density

$$h(z|x) = \frac{\exp\left(-\frac{\alpha\xi(x,z)}{2\Delta_{n,k}}\right)}{\int_{\mathcal{X}^k} \exp\left(-\frac{\alpha\xi(x,s)}{2\Delta_{n,k}}\right) ds} \quad (2)$$

where  $\alpha \geq 0$ ,  $z = (z_1, \dots, z_k)$  and  $x = (x_1, \dots, x_n)$ . In this case,  $Q_n$  has density  $h(z|x)$ . We'll discuss the exponential mechanism in more detail later.

There are many definitions of privacy but in this paper we focus on the following definition due to Dwork et al. (2006) and Dwork (2006).

**Definition 2.3.** *Let  $\alpha \geq 0$ . We say that  $Q_n$  satisfies  $\alpha$ -differential privacy if*

$$\sup_{\substack{x,y \in \mathcal{X}^n \\ \delta(x,y)=1}} \sup_{B \in \mathcal{B}} \frac{Q_n(B|X=x)}{Q_n(B|X=y)} \leq e^\alpha \quad (3)$$

where  $\mathcal{B}$  are the measurable sets on  $\mathcal{X}^k$ . The ratio is interpreted to be 1 whenever the numerator and denominator are both 0.

The definition of differential privacy is based on ratios of probabilities. It is crucial to measure closeness by ratios of probabilities since that protects rare cases which have small probability under  $Q_n$ . In particular, if changing one entry in the database  $X$  cannot change the probability distribution  $Q_n(\cdot|X=x)$  very much, then we can claim that a single individual cannot guess whether he is in the original database or not. The closer  $e^\alpha$  is to 1, the stronger privacy guarantee is. Thus, one typically chooses  $\alpha$  close to 0. See Dwork et al. (2006) for more discussion on these points. Indeed, suppose that two subjects each believe that one of them is in the original database. Given  $Z$  and full knowledge of  $P$  and  $Q_n$  can they test who is in  $X$ ? The answer is given in the following result. (In this result, we drop the assumption that the user does not know  $Q_n$ .)

**Theorem 2.4.** *Suppose that  $Z$  is obtained from a data release mechanism that satisfies  $\alpha$ -differential privacy. Any level  $\gamma$  test which is a function of  $Z$ ,  $P$  and  $Q_n$  of  $H_0 : X_i = s$  versus  $H_1 : X_i = t$  has power bounded above by  $\gamma e^\alpha$ .*

Thus, if  $Q_n$  satisfies differential privacy then it is virtually impossible to test the hypothesis that either of the two subjects is in the database since the power of such a test is nearly equal to its level. A similar calculation shows that if one does a Bayes test between  $H_0$  and  $H_1$  then the Bayes factor is always between  $e^{-2\alpha}$  and  $e^{2\alpha}$ . For more detail on the motivation for the definition as well as consequences, see Dwork et al. (2006), Dwork (2006), Ganta et al. (2008), Rastogi et al. (2009).

The following result, which is proved in McSherry and Talwar (2007) (Theorem 6), shows that the exponential mechanism always preserves differential privacy.

**Theorem 2.5.** (McSherry and Talwar, 2007) *The exponential mechanism satisfies the  $\alpha$ -differential privacy.*

To conclude this section we record a few useful facts. Let  $T(X, R)$  be a function of  $X$  and some auxiliary random variable  $R$  which is independent of  $X$ . After including this auxiliary random variable we define differential privacy as before. Specifically,  $T(X, R)$  satisfies differential privacy if for all  $B$ , and all  $x, x'$  with  $\delta(x, x') = 1$  we have that  $\mathbb{P}(T(X, R) \in B | X = x) \leq e^\alpha \mathbb{P}(T(X, R) \in B | X = x')$ . The third part is Proposition 1 from Dwork et al. (2006).

**Lemma 2.6.** *We have the following:*

1. *If  $T(X, R)$  satisfies differential privacy then  $U = h(T(X, R))$  also satisfies differential privacy for any measurable function  $h$ .*
2. *Suppose that  $g$  is a density function constructed from a random vector  $T(X, R)$  that satisfies differential privacy. Let  $Z = (Z_1, \dots, Z_k)$  be  $k$  iid draws from  $g$ . This defines a mechanism  $Q_n(B|X) = \mathbb{P}(Z \in B|X)$ . Then  $Q_n$  satisfies differential privacy for any  $k$ .*
3. *(Proposition 1 from Dwork et al. (2006).) Let  $f(x)$  be a function of  $x = (x_1, \dots, x_n)$  and define  $S(f) = \sup_{x, x': \delta(x, x')=1} \|f(x) - f(x')\|_1$  where  $\|a\|_1 = \sum_j |a_j|$ . Let  $R$  have density  $g(r) \propto e^{-\alpha|r|/S(f)}$ . Then  $T(X, R) = f(X) + R$  satisfies differential privacy.*

### 3 Informative Mechanisms

A challenge in privacy theory is to find  $Q_n$  that satisfies differential privacy and yet yields datasets  $Z$  that preserve information. Informally, a mechanism is informative if it is possible to make precise inferences from the released data  $Z_1, \dots, Z_k$ . Whether or not a mechanism is informative will depend on the goals of the inference. From a statistical perspective, we would like to infer  $P$  or functionals of  $P$  from  $Z$ . Blum et al. (2008) show that the probability content of some classes of intervals can be estimated accurately while preserving privacy. Their results motivated the current paper. We will assume throughout that the user has access to the sanitized data  $Z$  but not the mechanism  $Q_n$ . The question of how a data analyst can use knowledge of  $Q_n$  to improve inferences is left to future work.

There are many ways to measure the information in  $Z$ . One way is through distribution functions. Let  $F$  denote the cumulative distribution function (cdf) on  $\mathcal{X}$  corresponding to  $P$ . Thus  $F(x) = P(X \in (-\infty, x_1] \times \dots \times (-\infty, x_r])$  where  $x = (x_1, \dots, x_r)$ . Let  $\hat{F} \equiv \hat{F}_X$  denote the empirical distribution function corresponding to  $X$  and similarly let  $\hat{F}_Z$  denote the empirical distribution function corresponding to  $Z$ . Let  $\rho$  denote any distance measure on distribution functions.

**Definition 3.1.**  $Q_n$  is consistent with respect to  $\rho$  if  $\rho(F, \hat{F}_Z) \xrightarrow{P} 0$ .  $Q_n$  is  $\epsilon_n$ -informative if  $\rho(F, \hat{F}_Z) = O_P(\epsilon_n)$ .

An alternative to requiring  $\rho(F, \hat{F}_Z)$  to be small is to require  $\rho(\hat{F}, \hat{F}_Z)$  to be small. Or one could require  $Q_n(\rho(\hat{F}, \hat{F}_Z) > \epsilon | X = x)$  be small for all  $x$  as in Blum et al. (2008). These requirements are similar. Indeed, suppose  $\rho$  satisfies the triangle inequality and that  $\hat{F}$  is consistent in the  $\rho$  distance, that is,  $\rho(\hat{F}, F) \xrightarrow{P} 0$ . Assume further that  $\rho(\hat{F}, F) = O_P(\epsilon_n)$ . Then  $\rho(F, \hat{F}_Z) = O_P(\epsilon_n)$  implies that

$$\rho(\hat{F}, \hat{F}_Z) \leq \rho(\hat{F}, F) + \rho(F, \hat{F}_Z) = O_P(\epsilon_n);$$

Similarly,  $\rho(\hat{F}, \hat{F}_Z) = O_P(\epsilon_n)$  implies that  $\rho(F, \hat{F}_Z) = O_P(\epsilon_n)$ .

Let  $\mathbb{E}_{P, Q_n}$  denote the expectation under the joint distribution defined by  $P^n$  and  $Q_n$ . Sometimes we write  $\mathbb{E}$  when there is no ambiguity. Similarly, we use  $\mathbb{P}$  to denote the marginal probability

under  $P^n$  and  $Q_n$ :  $\mathbb{P}(A) = \int_A dQ_n(z_1, \dots, z_k | x_1, \dots, x_n) dP(x_1) \cdots dP(x_n)$  for  $A \in \mathcal{X}^k$ .

There are many possible choices for  $\rho$ . We shall mainly focus on the Kolmogorov-Smirnov (KS) distance  $\rho(F, G) = \sup_x |F(x) - G(x)|$  and the squared  $L_2$  distance  $\rho(F, G) = \int (f(x) - g(x))^2 dx$  where  $f = dF/d\mu$  and  $g = dG/d\mu$ . However, our results can be carried over to other distances as well.

Before proceeding let us note that we will need some assumptions on  $F$  otherwise we cannot have a consistent scheme as shown in the following theorem. The following result — essentially a re-expression of a result in Blum et al. (2008) in our framework — makes this clear.

**Theorem 3.2.** *Suppose that  $Q_n$  satisfies differential privacy and that  $\rho(F, G) = \sup_x |F(x) - G(x)|$ . Let  $F$  be a point mass distribution. Thus  $F(y) = I(y \geq x)$  for some point  $x \in [0, 1]$ . Then  $\widehat{F}_Z$  is inconsistent, that is, there is a  $\delta > 0$  such that  $\liminf_{n \rightarrow \infty} P^n(\rho(F, \widehat{F}_Z) > \delta) > 0$ .*

## 4 Sampling From a Histogram

The goal of this section is to give two concrete, simple data release methods that achieve differential privacy. The idea is to draw a random sample from histogram. The first scheme draws observations from a smoothed histogram. The second scheme draws observations from a randomly perturbed histogram. We use the histogram for its familiarity and simplicity and because it is used in applications of differential privacy. We will see that the histogram has to be carefully constructed to ensure differential privacy. We then compare the two schemes by studying the accuracy of the inferences from the released data. We will see that the accuracy depends both on how the histogram is constructed and on what measure of accuracy we use.

Let  $L > 0$  be a constant and suppose that  $p = dP/d\mu \in \mathcal{P}$  where

$$\mathcal{P} = \left\{ p : |p(x) - p(y)| \leq L|x - y| \right\} \quad (4)$$

is the class of Lipschitz functions. We assume throughout this section that  $p \in \mathcal{P}$ . The minimax rate of convergence for density estimators in squared  $L_2$  distance for  $\mathcal{P}$  is  $n^{-2/(2+r)}$  (Scott, 1992).



Let  $h = h_n$  be a binwidth such that  $0 < h < 1$  and such that  $m = 1/h^r$  is an integer. Partition  $\mathcal{X}$  into  $m$  bins  $\{B_1, \dots, B_m\}$  where each bin  $B_j$  is a cube with sides of length  $h$ . Let  $I(\cdot)$  denote the indicator function. Let  $\hat{f}_m$  denote the corresponding histogram estimator on  $\mathcal{X}$ , namely,

$$\hat{f}_m(x) = \sum_{j=1}^m \frac{\hat{p}_j}{h^r} I(x \in B_j)$$

where  $\hat{p}_j = C_j/n$  and  $C_j = \sum_{i=1}^n I(X_i \in B_j)$  is the number of observations in  $B_j$ . Recall that  $\hat{f}_m$  is a consistent estimator of  $p$  if  $h = h_n \rightarrow 0$  and  $nh_n^r \rightarrow \infty$ . Also, the optimal choice of  $m = m_n$  for  $L_2$  error under  $\mathcal{P}$  is  $m_n \asymp n^{r/(2+r)}$ , in which case  $\int (p - \hat{f}_m)^2 = O_P(n^{-2/(2+r)})$  (Scott, 1992). Here,  $a_n \asymp b_n$  means that both  $a_n/b_n$  and  $b_n/a_n$  are bounded for large  $n$ .

## 4.1 Sampling from a Smoothed Histogram

The first method for generating released data  $Z$  from a histogram while achieving differential privacy proceeds as follows. Recall that the sample space is  $[0, 1]^r$ . Fix a constant  $0 < \delta < 1$  and define the smoothed histogram

$$\hat{f}_{m,\delta}(x) = (1 - \delta)\hat{f}_m(x) + \delta. \quad (5)$$

**Theorem 4.1.** *Let  $Z = (Z_1, \dots, Z_k)$  where  $Z_1, \dots, Z_k$  are  $k$  iid draws from  $\hat{f}_{m,\delta}(x)$ . If*

$$k \log \left( \frac{(1 - \delta)m}{n\delta} + 1 \right) \leq \alpha \quad (6)$$

*then  $\alpha$ -differential privacy holds.*

Note that for  $\delta \rightarrow 0$  and  $\frac{m}{n\delta} \rightarrow 0$ ,  $\log \left( \frac{(1 - \delta)m}{n\delta} + 1 \right) = \frac{m}{n\delta}(1 + o(1)) \approx \frac{m}{n\delta}$ . Thus (6) is approximately the same as requiring

$$\frac{mk}{\delta} \leq n\alpha. \quad (7)$$

Equation (7) shows an interesting tradeoff between  $m$ ,  $k$  and  $\delta$ . We note that sampling from the usual histogram corresponding to  $\delta = 0$  does not preserve differential privacy.

Now we consider how to choose  $m, k, \delta$  to minimize  $\mathbb{E}(\rho(F, \widehat{F}_Z))$  while satisfying (6). Here,  $\mathbb{E}$  is the expectation under the randomness due to sampling from  $P$  and due to the privacy mechanism  $Q_n$ . Thus, for any measurable function  $h$ ,

$$\mathbb{E}(h(Z)) = \int \int h(z_1, \dots, z_k) dQ_n(z_1, \dots, z_k | x_1, \dots, x_n) dP(x_1) \cdots dP(x_n).$$

Now we give a result that shows how accurate the inferences are in the KS distance using the smoothed histogram sampling scheme.

**Theorem 4.2.** *Suppose that  $Z_1, \dots, Z_k$  are drawn as described in the previous theorem. Suppose (4) holds. Let  $\rho$  be the KS distance. Then choosing  $m \asymp n^{r/(6+r)}$ ,  $k \asymp m^{4/r} = n^{4/(6+r)}$  and  $\delta = (mk/n\alpha)$  minimizes  $\mathbb{E}\rho(F, \widehat{F}_Z)$  subject to (6). In this case,  $\mathbb{E}\rho(F, \widehat{F}_Z) = O\left(\frac{\sqrt{\log n}}{n^{2/(6+r)}}\right)$ .*

In this case we see that we have consistency since  $\rho(F, \widehat{F}_Z) = o_P(1)$  but the rate is slower than the minimax rate of convergence for density estimators in KS distance, which is  $n^{-1/2}$ . Now let  $\widehat{q}_j = \#\{Z_i \in B_j\}/k$  and

$$\rho(F, \widehat{F}_Z) = \int (p(x) - \widehat{f}_Z(x))^2 dx, \text{ where } \widehat{f}_Z(x) = h^{-r} \sum_{j=1}^m \widehat{q}_j I(x \in B_j). \quad (8)$$

**Theorem 4.3.** *Assume the conditions of the previous theorem. Let  $\rho$  be the squared  $L_2$  distance as defined in (8). Then choosing*

$$m \asymp n^{r/(2r+3)}, \quad k \asymp n^{(r+2)/(2r+3)}, \quad \delta \asymp n^{-1/(r+3)}$$

*minimizes  $\mathbb{E}\rho(F, \widehat{F}_Z)$  subject to (6). In this case,  $\mathbb{E}\rho(F, \widehat{F}_Z) = O(n^{-2/(2r+3)})$ .*

Again, we have consistency but the rate is slower than the minimax rate which is  $n^{-2/(2+r)}$ . (Scott, 1992)

## 4.2 Sampling From a Perturbed Histogram

The second method, which we call the sampling from a perturbed histogram, is due to Dwork et al. (2006). Recall that  $C_j$  is the number of observations in bin  $B_j$ . Let  $D_j = C_j + \nu_j$  where  $\nu_1, \dots, \nu_m$  are independent, identically distributed draws from a Laplace density with mean 0 and variance  $8/\alpha^2$ . Thus the density of  $\nu_j$  is  $g(\nu) = (\alpha/4)e^{-|\nu|\alpha/2}$ . Dwork et al. (2006) show that releasing  $D = (D_1, \dots, D_m)$  preserves differential privacy. However, our goal is to release a database  $Z = (Z_1, \dots, Z_k)$  rather than just a set of counts. Now define

$$\tilde{D}_j = \max\{D_j, 0\} \quad \text{and} \quad \hat{q}_j = \tilde{D}_j / \sum_s \tilde{D}_s.$$

Since  $D$  preserves differential privacy, it follows from Lemma 2.6 that  $(\hat{q}_1, \dots, \hat{q}_m)$  also preserve differential privacy; Moreover, any sample  $Z = (Z_1, \dots, Z_k)$  from  $\tilde{f}(x) = h^{-r} \sum_{j=1}^m \hat{q}_j I(x \in B_j)$  preserve differential privacy for any  $k$ .

**Theorem 4.4.** *Let  $Z = (Z_1, \dots, Z_k)$  be drawn from  $\tilde{f}(x) = h^{-r} \sum_{j=1}^m \hat{q}_j I(x \in B_j)$ . Assume that there exists a constant  $1 \leq C < \infty$  such that  $\sup_x p(x) = C$ .*

(1) *Let  $\rho$  be the  $L_2$  distance and  $\hat{f}_Z$  be as defined in (8). Let  $m \asymp n^{r/(2+r)}$  and let  $k \geq n$ . Then we have  $\mathbb{E}\rho(F, \hat{F}_Z) = O(n^{-2/(2+r)})$ .*

(2) *Let  $\rho$  be the KS distance. Let  $m \asymp n^{r/(2+r)}$ . Then  $\mathbb{E}\rho(F, \hat{F}_Z) = O\left(\min\left(\frac{\log n}{n^{2/(2+r)}}, \sqrt{\frac{\log n}{n}}\right)\right)$ .*

Hence, this method achieves the minimax rate of convergence in  $L_2$  while the first data release method does not. This suggests that the perturbation method is preferable for the  $L_2$  distance. The perturbation method does not achieve the minimax rate of convergence in KS distance; in fact, the exponential mechanism based method achieves a better rate as we shown in Section 5 (Theorem 5.4). We examine this method numerically in Section 7.

Another approach to histograms is given by Machanavajjhala et al. (2008). They put a Dirichlet  $(a_1, \dots, a_m)$  prior on the cell probabilities  $p_1, \dots, p_m$  where  $p_j = \mathbb{P}(X_i \in B_j)$ . The corresponding posterior is Dirichlet  $(a_1 + C_1, \dots, a_m + C_m)$ . Next they draw  $q = (q_1, \dots, q_m)$  from the posterior and finally they sample new cell counts  $D = (D_1, \dots, D_m)$  from a Multinomial  $(k, q)$ . Thus, the

distribution of  $D$  given  $X$  is

$$\mathbb{P}(D = d|X) = \frac{\prod_{j=1}^m \Gamma(d_j + a_j + C_j)}{\Gamma(k + n + \sum_j a_j)}.$$

They show that differential privacy requires  $a_j + C_j \geq k/(e^\alpha - 1)$  for all  $j$ . If we take  $a_1 = a_2 = \dots = a_m$  then this is similar to the first histogram-based data release method we discussed in this section. They also suggest a weakened version of differential privacy.

## 5 Exponential Mechanism

In this section we will consider the exponential mechanism in some detail. We'll derive some general results about accuracy and apply the method to the mean, and to density estimation. Specifically, we will show the following for exponential mechanisms:

1. Choosing the size  $k$  of the released database is delicate. Taking  $k$  too large compromises privacy. Taking  $k$  too small compromises accuracy.
2. The accuracy of the exponential scheme can be bounded by a simple formula. This formula has a term that measures how likely it is for a distribution based on sample size  $k$ , to be in a small ball around the true distribution. In probability theory, this is known as a small ball probability.
3. The formula can be applied to several examples such as the KS distance, the mean, and nonparametric density estimation using orthogonal series. In each case we can use our results to choose  $k$  and to find the rate of convergence of an estimator based on the sanitized data.

In light of Theorem 3.2, we know that some assumptions are needed on  $P$ . We shall assume throughout this section that  $P$  has a bounded density  $p$ ; note that this is a weaker condition than (4).

Recall the exponential mechanism. We draw the vector  $Z = (Z_1, \dots, Z_k)$  from  $h(z|x)$  where

$$h(z|x) = \frac{g_x(z)}{\int_{[0,1]^k} g_x(s) ds}, \quad \text{where } g_x(z) = \exp\left(-\frac{\alpha \rho(\widehat{F}_x, \widehat{F}_z)}{2\Delta_{n,k}}\right) \quad \text{and} \quad (9)$$

$$\Delta \equiv \Delta_{n,k} = \sup_{\substack{x, y \in \mathcal{X}^n \\ \delta(x, y) = 1}} \sup_{z \in \mathcal{X}^k} |\rho(\widehat{F}_x, \widehat{F}_z) - \rho(\widehat{F}_y, \widehat{F}_z)|.$$

**Lemma 5.1.** For KS distance  $\Delta_{n,k} \leq \frac{1}{n}$ .

This framework is used in Blum et al. (2008). For the rest of this section, assume that  $Z = (Z_1, \dots, Z_k)$  are drawn from an exponential mechanism  $Q_n$ .

**Definition 5.2.** Let  $F$  denote the cumulative distribution function on  $\mathcal{X}$  corresponding to  $P$ . Let  $\widehat{G}$  denote the empirical cdf from a sample of size  $k$  from  $P$ , and let

$$R(k, \epsilon) = P^k(\rho(F, \widehat{G}) \leq \epsilon).$$

$R(k, \epsilon)$  is called the small ball probability associated with  $\rho$ .

The following theorem bounds the accuracy of the estimator from the sanitized data by a simple formula involving the small ball probability.

**Theorem 5.3.** Assume that  $P$  has a bounded density  $p$ , and that there exists  $\epsilon_n \rightarrow 0$  such that

$$\mathbb{P}\left(\rho(F, \widehat{F}_X) > \frac{\epsilon_n}{16}\right) = O\left(\frac{1}{n^c}\right) \quad (10)$$

for some  $c > 1$ . Further suppose that  $\rho$  satisfies the triangle inequality. Let  $Z = (Z_1, \dots, Z_k)$  be drawn from  $g_x(z)$  given in (9). Then,

$$\mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n\right) \leq \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\epsilon_n}{16\Delta}\right)}{R(k, \epsilon_n/2)} + O\left(\frac{1}{n^c}\right). \quad (11)$$

Thus, if we can choose  $k = k_n$  in such a way that the right hand side of (11) goes to 0, then the mechanism is consistent. We now show some examples that satisfy these conditions and we show how to choose  $k_n$ .

## 5.1 The KS Distance

**Theorem 5.4.** *Suppose that  $P$  has a bounded density  $p$  and let  $B := \log \sup_x p(x) > 0$ . Let  $Z = (Z_1, \dots, Z_k)$  be drawn from  $g_x(z)$  given in (9) with  $\rho$  being the KS distance. By requiring that  $k_n \asymp \left(\frac{3\alpha}{B}\right)^{2/3} n^{2/3}$ , we have for  $\epsilon_n = 2 \left(\frac{B}{3\alpha}\right)^{1/3} n^{-1/3}$ , and for  $\rho$  being the KS distance,*

$$\rho(F, \widehat{F}_Z) = O_P(\epsilon_n). \quad (12)$$

Note that  $\rho(F, \widehat{F}_Z)$  converges to 0 at a slower rate than  $\rho(F, \widehat{F}_X)$ . We thus see that the rate after sanitization is  $n^{-1/3}$  which is slower than the optimal rate of  $n^{-1/2}$ . It is an open question whether this rate can be improved.

## 5.2 The Mean

It is interesting to consider what happens when  $\rho(F, \widehat{F}_Z) = \|\mu - \bar{Z}\|^2$  where  $\mu = \int x dP(x)$  and  $\bar{Z}$  is the sample mean of  $Z$ . In this case  $\Delta \leq r/n$ . Thus,  $h(u|x) \approx e^{-n\|\bar{X} - \bar{Z}\|^2/(2\alpha)}$  so, approximately,  $Z_1, \dots, Z_k \sim N(\bar{X}, k\alpha/n)$ . Indeed, it suffices to take  $k = 1$  in this case since then  $\bar{Z} = \bar{X} + O_P(1/\sqrt{n})$ . Thus  $\bar{Z}$  converges at the same rate as  $\bar{X}$ . This is not surprising: preserving a single piece of information requires a database of size  $k = 1$ .

## 6 Orthogonal Series Density Estimation

In this section, we develop an exponential scheme based on density estimation and we compare it to the perturbation approach. For simplicity we take  $r = 1$ . Let  $\{1, \psi_1, \psi_2, \dots\}$  be an orthonormal basis for  $L_2(0, 1) = \{f : \int_0^1 f^2(x) dx < \infty\}$  and assume that  $p \in L_2(0, 1)$ . Hence

$$p(x) = 1 + \sum_{j=1}^{\infty} \beta_j \psi_j(x) \quad \text{where} \quad \beta_j = \int_0^1 \psi_j(x) p(x) dx.$$

We assume that the basis functions are uniformly bounded so that

$$c_0 \equiv \sup_j \sup_x |\psi_j(x)| < \infty. \quad (13)$$

Let  $\mathcal{B}(\gamma, C)$  denote the Sobolev ellipsoid

$$\mathcal{B}(\gamma, C) = \left\{ \beta = (\beta_1, \beta_2, \dots) : \sum_{j=1}^{\infty} \beta_j^2 j^{2\gamma} \leq C^2 \right\}$$

where  $\gamma > 1/2$ . Let

$$\mathcal{P}(\gamma, C) = \left\{ p(x) = 1 + \sum_{j=1}^{\infty} \beta_j \psi_j(x) : \beta \in \mathcal{B}(\gamma, C) \right\}.$$

The minimax rate of convergence in  $L_2$  norm for  $\mathcal{P}(\gamma, C)$  is  $n^{-2\gamma/(2\gamma+1)}$  (Efremovich, 1999). Thus

$$\inf_{\hat{p}} \sup_{P \in \mathcal{P}(\gamma, C)} E \int (\hat{p}(x) - p(x))^2 dx \geq c_1 n^{-2\gamma/(2\gamma+1)}$$

for some  $c_1 > 0$ . This rate is achieved by the estimator

$$\hat{p}(x) = 1 + \sum_{j=1}^{m_n} \hat{\beta}_j \psi_j(x) \quad (14)$$

where  $m_n = n^{1/(2\gamma+1)}$  and  $\hat{\beta}_j = n^{-1} \sum_{i=1}^n \psi_j(X_i)$ . See Efremovich (1999).

For a function  $u \in L_2(0, 1)$ , let us define  $\|u\|_{\ell_2} = \left( \int_0^1 |u(x)|^2 dx \right)^{1/2}$ , which is a norm on  $L_2(0, 1)$ . Now consider an exponential mechanism based on

$$\xi(X, Z) = \left( \int (\hat{p}(x) - \hat{p}^*(x))^2 dx \right)^{1/2} := \|\hat{p} - \hat{p}^*\|_{\ell_2} \quad \text{where} \quad (15)$$

$$\hat{p}^*(x) = 1 + \sum_{j=1}^{m_k} \hat{\beta}_j^* \psi_j(x), \quad \text{for } m_k = k^{\frac{1}{2\gamma+1}} \text{ and } \hat{\beta}_j^* = k^{-1} \sum_{i=1}^k \psi_j(Z_i). \quad (16)$$

**Lemma 6.1.** Under the above scheme we have  $\Delta \leq \frac{2c_0^2 m_n}{n}$  for  $c_0$  as defined in (13). Hence,

$$g(z|x) = \exp\left(-\frac{\alpha \|\widehat{p}^* - \widehat{p}\|_{\ell_2}}{\Delta}\right) \leq \exp\left(-\frac{\alpha n \|\widehat{p}^* - \widehat{p}\|_{\ell_2}}{2c_0^2 m_n}\right) \text{ almost surely.} \quad (17)$$

**Theorem 6.2.** Let  $Z = (Z_1, \dots, Z_k)$  be drawn from  $g_x(z)$  given in (17). Assume that  $\gamma > 1$ . If we choose  $k \asymp \sqrt{n}$  then

$$\rho^2(p, \widehat{p}^*) = O_P\left(n^{-\frac{\gamma}{2\gamma+1}}\right).$$

We conclude that the sanitized estimator converges at a slower rate than the minimax rate. Now we compare this to the perturbation approach. Let  $Z = (Z_1, \dots, Z_k)$  be an iid sample from

$$\widehat{q}(x) = 1 + \sum_{j=1}^{m_n} (\widehat{\beta}_j + \nu_j) \psi_j(x)$$

where  $\nu_1, \dots, \nu_m$  are iid draws from a Laplace distribution with density  $g(\nu) = (n\alpha/(2c_0 m))e^{-n\alpha|\nu|/(c_0 m)}$ .

Thus, in the notation of 2.6,  $R = (\nu_1, \dots, \nu_m)$ . It follows from Lemma 2.6 that, for any  $k$ , this preserves differential privacy. If  $\widehat{q}(x) < 0$  for any  $x$  then we replace  $\widehat{q}$  by  $\widehat{q}(x)I(\widehat{q}(x) > 0) / \int \widehat{q}(s)I(\widehat{q}(s) > 0)ds$  as in Hall and Murison (1993).

**Theorem 6.3.** Let  $Z = (Z_1, \dots, Z_k)$  be drawn from  $\widehat{q}$ . Assume that  $\gamma > 1$ . If we choose  $k \geq n$ , then

$$\rho^2(p, \widehat{p}_Z) = O_P\left(n^{-\frac{2\gamma}{2\gamma+1}}\right)$$

where  $\widehat{p}_Z$  is the orthogonal series density estimator based on  $Z$ .

Hence, again, the perturbation technique achieves the minimax rate of convergence and so appears to be superior to the exponential mechanism. We do not know if this is because the exponential mechanism is inherently less accurate, or if our bounds for the exponential mechanism are not tight enough.



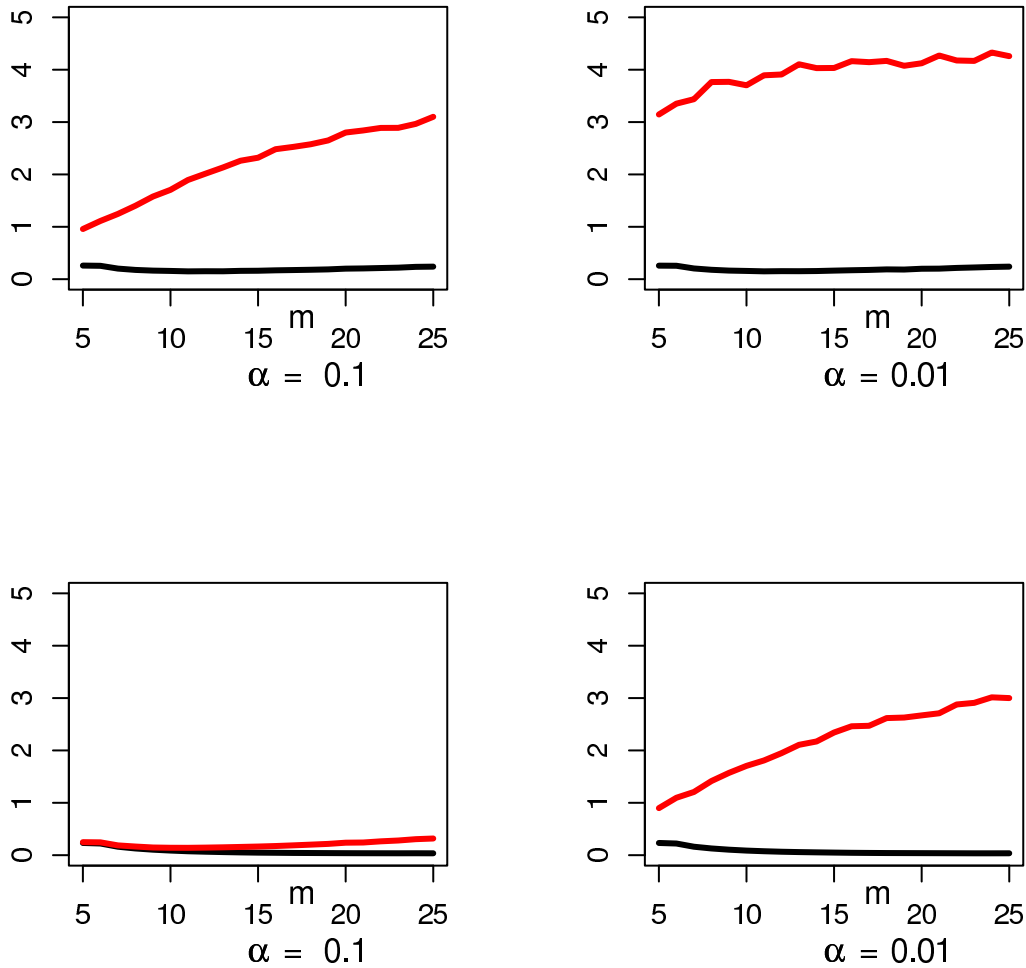


Figure 1: Top two plots  $n = 100$ . Bottom two plots  $n = 1,000$ . Each plot shows the mean integrated squared error of the histogram. The lower line is from the histogram based on the original data. The upper line is based on the perturbed histogram.

## 7 Example

Here we consider a small simulation study to see the effect of perturbation on accuracy. We focus on the histogram perturbation method with  $r = 1$ . We take the true density of  $X$  to be a Beta(10,10) density. We considered sample sizes  $n = 100$  and  $n = 1,000$  and privacy levels  $\alpha = 0.1$ , and  $\alpha = 0.01$ . We take  $\rho$  to be squared error distance. Figure 1 shows the results of 1,000 simulations for various numbers of bins  $m$ .

As expected, smaller values of  $\alpha$  induce a larger information loss which manifests itself as a

larger mean squared error. Despite the fact that the perturbed histogram achieves the minimax rate, the error is substantially inflated by the perturbation. This means that the constants in the risk are important, not just the rate. Also, the risk of the sanitized histograms is much more sensitive to the choice of the number of cells than the original histogram is.

We repeated the simulations with a bimodal density, namely,  $p(x)$  being an equal mixture of a Beta(10,3) density and Beta(3,10) density. The results turned out to be nearly identical to those above.

## 8 Conclusion

Differential privacy is an important type of privacy guarantee when releasing data. Our goal has been to present the idea in statistical language and then to show that loss functions based on distributions and densities can be useful for comparing privacy mechanisms.

We have seen that sampling from a histogram leads to differential privacy as long as either the histogram is shifted away from 0 by a factor  $\delta$  or if the cells are perturbed appropriately. The latter method achieves a faster rate of convergence in  $L_2$  distance. But, the simulation showed that the risk can nonetheless be quite large. This suggests that more work is needed to get precise finite sample risk bounds. Also, the choice of the smoothing parameter (number of cells in the histogram) has a larger effect on the sanitized histogram than on the original histogram.

We also studied the exponential mechanism. Here we derived a formula for assessing the accuracy of the method. The formula involves small ball probabilities. As far as we know, the connection between differential privacy and small ball probabilities has not been observed before.

Minimaxity is desirable for any statistical procedure. We have seen that in some cases the minimax rate is achieved and in some cases it is not. We do not yet have a complete minimax theory for differential privacy and this is the focus of our current work. We close with some open questions.

1. When is it possible for  $\rho(F, \widehat{F}_Z)$  to have the same rate as  $\rho(F, \widehat{F}_X)$ ?
2. When adaptive minimax methods are used, such as adapting to  $\gamma$  in Section 6 or when using

wavelet estimation methods, is some form of adaptivity preserved after sanitization?

3. Many statistical methods involve some sort of risk minimization. A example is choosing a bandwidth by cross-validation. What is the effect of sanitization on these procedures?
4. Are there other, better methods of sanitization that preserve differential privacy?

## 9 Proofs

### 9.1 Proof of Theorem 2.4

Without loss of generality take  $i = 1$ . Let  $M_0(B) = \int Q(B|s, x_2, \dots, x_n) dP(x_2, \dots, x_n)$  and  $M_1(B) = \int Q(B|t, x_2, \dots, x_n) dP(x_2, \dots, x_n)$ . By the Neyman-Pearson lemma, the highest power test is to reject  $H_0$  when  $U > u$  where  $U(z) = (dM_1/dM_0)(z)$  and  $u$  is chosen so that  $\int I(U(z) > u) dM_0(z) \leq \gamma$ . Since  $(s, x_2, \dots, x_n)$  and  $(t, x_2, \dots, x_n)$  differ in only one coordinate,  $M_1(B) \leq e^\alpha M_0(B)$  and so the power is  $M_1(U > u) \leq e^\alpha M_0(U > u) \leq \gamma e^\alpha$ .  $\square$

### 9.2 Proof of Lemma 2.6

For the first part simply note that  $\mathbb{P}(h(T(X, R)) \in B|X = x) = \mathbb{P}(T(X, R) \in h^{-1}(B)|X = x) \leq e^\alpha \mathbb{P}(T(X, R) \in h^{-1}(B)|X = x') = e^\alpha \mathbb{P}(h(T(X, R)) \in B|X = x')$ .

For the second part, let  $Z = (Z_1, \dots, Z_k)$  and note that  $Z$  is independent of  $X$  given  $T(X, R)$ . Let  $H$  be the distribution of  $T(X, R)$ . Hence,

$$\begin{aligned}
 \mathbb{P}(Z \in B|X = x) &= \int \mathbb{P}(Z \in B|X = x, T = t) dH(t|X = x) dt \\
 &= \int \mathbb{P}(Z \in B|T = t) dH(t|X = x) dt \\
 &= \int \mathbb{P}(Z \in B|T = t) \frac{dH(t|X = x)}{dH(t|X = x')} dH(t|X = x') \\
 &\leq e^\alpha \int \mathbb{P}(Z \in B|T = t) dH(t|X = x') \\
 &= e^\alpha \mathbb{P}(Z \in B|X = x').
 \end{aligned}$$

### 9.3 Proof of Theorem 3.2

Our proof is adapted from an argument given in Theorem 5.1. of Blum et al. (2008). Let  $r = 1$  so that  $\mathcal{X} = [0, 1]$ . Let  $P = \delta_0$  where  $\delta_0$  denotes a point mass at 0. Then  $P^n(X = X_{(0)}) = 1$  where  $X_{(0)} \equiv \{0, \dots, 0\}$ . Assume that  $Q_n$  is consistent. Since  $F(0) = 1$ , it follows that for any  $\delta > 0$ ,  $\mathbb{P}(\widehat{F}_Z(0) > 1 - \delta) \rightarrow 1$ . But since  $\mathbb{P}(\cdot) = \mathbb{E}_P Q_n(\cdot|X)$  and since  $P^n(X = X_{(0)}) = 1$ , this implies that  $Q_n(\widehat{F}_Z(0) > 1 - \delta|X = X_{(0)}) \rightarrow 1$ .

Let  $v > 0$  be any point in  $[0, 1]$  such that  $Q_n(Z = v|X = X_{(0)}) = 0$ . Let  $X_{(1)} = \{v, 0, \dots, 0\}$ ,  $X_{(2)} = \{v, v, 0, \dots, 0\}$ ,  $\dots$ ,  $X_{(n)} = \{v, v, \dots, v\}$ . By assumption,  $Q_n(Z = X_{(j)}|X = X_{(0)}) = 0$  for all  $j \geq 1$ . Differential privacy implies that  $Q_n(Z = X_{(j)}|X = X_{(1)}) = 0$  for all  $j \geq 1$ . Applying differential privacy again implies that  $Q_n(Z = X_{(j)}|X = X_{(2)}) = 0$  for all  $j \geq 1$ . Continuing this way, we conclude that  $Q_n(Z = X_{(j)}|X = X_{(n)}) = 0$  for all  $j \geq 1$ .

Next let  $P = \delta_v$ . Arguing as before, we know that  $Q_n(\widehat{F}_Z(v) < 1 - \delta|X = X_{(n)}) \rightarrow 0$ . And since  $F(v-) = 0$  we also have that  $Q_n(\widehat{F}_Z(v-) > \delta|X = X_{(n)}) \rightarrow 0$ . Here,  $F(v-) = \lim_{i \rightarrow \infty} F(v_i)$  where  $v_1 < v_2 < \dots$  and  $v_i \rightarrow v$ . Hence, for  $j/n > 1 - \delta$ ,  $Q_n(Z = X_{(j)}|X = X_{(n)}) > 0$  which is a contradiction.  $\square$

### 9.4 Proof of Theorem 4.1

Suppose that  $X$  differs from  $Y$  in at most one observation. Let  $\widehat{f}$  denote the perturbed histogram  $\widehat{f}_{m,\delta}$  based on  $X$  and let  $\widehat{g}_{m,\delta}$  denote the histogram based on  $Y$ , such that  $X$  and  $Y$  differ in one entry. We also use  $\widehat{p}_j(X)$  and  $\widehat{p}_j(Y)$  for cell proportions. Note that  $|\widehat{p}_j(X) - \widehat{p}_j(Y)| < 1/n$  by definition. It is clear that the maximum density ratio for a single draw  $x_i$ , or all  $i$ , occurs in one bin  $B_j$ . Now consider  $\mathbf{x} = (x_1, \dots, x_i)$  such that for all  $i = 1, \dots, k$ , we have  $x_i \in B_j \subset [0, 1]^r$  and the following bounds.

1. Let  $\widehat{p}_j(Y) = 0$ ; then in order to maximize  $\widehat{f}(\mathbf{x})/\widehat{g}(\mathbf{x})$ , we let  $\widehat{p}_j(X) = 1/n$  and obtain

$$\frac{\widehat{f}(\mathbf{x})}{\widehat{g}(\mathbf{x})} = \prod_{i=1}^k \frac{\widehat{f}_{m,\delta}(x_i)}{\widehat{g}_{m,\delta}(x_i)} \leq \left( \frac{(1-\delta)m(1/n) + \delta}{\delta} \right)^k = \left( \frac{(1-\delta)m}{n\delta} + 1 \right)^k ;$$

2. Otherwise, we let  $\widehat{p}_j(Y) \geq 1/n$ , (as by definition of  $\widehat{p}_j$ , it takes  $z/n$  for non-negative integers  $z$ ) and let  $\widehat{p}_j(X) = \widehat{p}_j(Y) \pm 1/n$ . Now it is clear that in order to maximize the density ratio at  $x$ , we may need to reverse the role of  $X$  and  $Y$ ,

$$\begin{aligned} \max \left( \frac{\widehat{g}(\mathbf{x})}{\widehat{f}(\mathbf{x})}, \frac{\widehat{f}(\mathbf{x})}{\widehat{g}(\mathbf{x})} \right) &\leq \max \left( \left( \frac{(1-\delta)m\widehat{p}_j + \delta}{(1-\delta)m(\widehat{p}_j - (1/n)) + \delta} \right)^k, \left( \frac{(1-\delta)m(\widehat{p}_j + 1/n) + \delta}{(1-\delta)m\widehat{p}_j + \delta} \right)^k \right), \\ &\leq \max \left( \frac{(1-\delta)m(1/n)}{(1-\delta)m(\widehat{p}_j - (1/n)) + \delta} + 1 \right)^k \\ &\leq \left( \frac{(1-\delta)m}{n\delta} + 1 \right)^k, \end{aligned}$$

where the maximum is achieved when  $\widehat{p}_j(Y) = 1/n$  and  $\widehat{p}_j(X) = 0$ , given a fixed set of parameters  $m, n, \delta$ .

Thus we have

$$\sup_{\mathbf{x} \in ([0,1]^r, \dots, [0,1]^r)} \frac{\widehat{f}(\mathbf{x})}{\widehat{g}(\mathbf{x})} \leq \left( \frac{(1-\delta)m}{n\delta} + 1 \right)^k,$$

and the theorem holds.  $\square$

## 9.5 Proof of Theorem 4.2

Recall that  $\widehat{F}_Z$  denotes the empirical distribution function corresponding to  $Z = (Z_1, \dots, Z_k)$ , where  $Z_i \in [0, 1]^r$  for all  $i$  are i.i.d. draws from density function  $\widehat{f}_{m,\delta}(x)$  as in (5) given  $X = (X_1, \dots, X_n)$ . Let  $U$  denote the uniform cdf on  $[0, 1]^r$ . Given  $X = (X_1, \dots, X_n)$  drawn from a distribution whose cdf is  $F$ , let  $\widehat{f}_m$  denote the histogram estimator on  $X$  and let  $\widehat{F}_m(x) = \int_0^x \widehat{f}_m(s) ds$  and  $\widehat{F}_{m,\delta}(x) = (1-\delta)\widehat{F}_m(x) + \delta U(x)$ . Define  $F_m(x) = \mathbb{E}(\widehat{F}_m(x))$  and  $\bar{f}_m(x) = \mathbb{E}(\widehat{f}_m(x))$ .

The Vapnik-Chervonenkis dimension of the class of sets of the form  $\{(-\infty, x_1] \times \dots \times (-\infty, x_r]\}$  is  $r$  and so by the standard Vapnik-Chervonenkis bound, we have for  $\epsilon > 0$  that

$$\mathbb{P} \left( \sup_{t \in [0,1]^r} |\widehat{F}_X(t) - F(t)| > \epsilon \right) \leq 8n^r \exp \left\{ -\frac{n\epsilon^2}{32} \right\} \leq \exp \left\{ -\frac{n\epsilon^2}{64} \right\} \quad (18)$$

for large  $n$ . Hence,  $\mathbb{E} \sup_{t \in [0,1]^r} |\widehat{F}_X(t) - F(t)| = O \left( \sqrt{\frac{r \log n}{n}} \right)$ . Given  $X$ , we have  $Z_1, \dots, Z_k \sim$

$\widehat{F}_{m,\delta}$  and so  $\mathbb{E} \sup_{[0,1]^r} |\widehat{F}_Z(t) - \widehat{F}_{m,\delta}(t)| = O\left(\sqrt{\frac{r \log k}{k}}\right)$ . Thus,

$$\begin{aligned} \mathbb{E} \sup_{x \in [0,1]^r} \left| \widehat{F}_Z(x) - F(x) \right| &\leq \mathbb{E} \sup_x |\widehat{F}_Z(x) - \widehat{F}_{m,\delta}(x)| + \mathbb{E} \sup_x |\widehat{F}_{m,\delta}(x) - F(x)| \\ &\leq \mathbb{E} \sup_x |\widehat{F}_Z(x) - \widehat{F}_{m,\delta}(x)| + \mathbb{E} \sup_x |\widehat{F}_m(x) - F(x)| + \delta \\ &\leq \mathbb{E} \sup_x |\widehat{F}_Z(x) - \widehat{F}_{m,\delta}(x)| + \mathbb{E} \sup_x |\widehat{F}_m(x) - F(x)| + \delta \\ &= O\left(\sqrt{\frac{r \log k}{k}}\right) + \mathbb{E} \sup_x |\widehat{F}_m(x) - F(x)| + \delta. \end{aligned}$$

By the triangle inequality, we have for all  $x \in [0, 1]^r$ ,

$$\left| \widehat{F}_m(x) - F(x) \right| \leq \left| \widehat{F}_m(x) - F_m(x) \right| + |F_m(x) - F(x)|,$$

and hence

$$\begin{aligned} \mathbb{E} \sup_{x \in [0,1]^r} \left| \widehat{F}_m(x) - F(x) \right| &\leq \mathbb{E} \sup_{x \in [0,1]^r} \left| \widehat{F}_m(x) - F_m(x) \right| + \mathbb{E} \sup_{x \in [0,1]^r} |F_m(x) - F(x)| \\ &= O\left(\sqrt{\frac{r \log n}{n}}\right) + \mathbb{E} \sup_{x \in [0,1]^r} |F_m(x) - F(x)| \end{aligned} \quad (19)$$

where the last step follows from the VC bound as in (18) for  $F_m(x)$ .

Next we bound  $\sup_{x \in [0,1]^r} |F_m(x) - F(x)|$ . Now  $F(x) = P(A)$  where  $A = \{(s_1, \dots, s_r) : s_i \leq x_i, i = 1, \dots, r\}$ . If  $x = (j_1 h, \dots, j_r h)$  for some integers  $j_1, \dots, j_r$  then  $F(x) - F_m(x) = 0$ . For  $x$  not of this form, let  $\tilde{x} = (j_1 h, \dots, j_r h)$  where  $j_i = \lfloor x_i/h \rfloor$ . Let  $R = \{(s_1, \dots, s_r) : s_i \leq \tilde{x}_i, i = 1, \dots, r\}$ . So

$$\begin{aligned} F(x) - F_m(x) &= P(A) - P_m(A) = P(R) - P_m(R) + P(A \setminus R) - P_m(A \setminus R) \\ &= P(A \setminus R) - P_m(A \setminus R) \end{aligned} \quad (20)$$

where  $P_m(B) = \int_B dF_m(u)$  and the set  $A \setminus R$  intersects at most  $rh/h^r$  number of cubes in  $\{B_1, \dots, B_m\}$ , given that  $\text{Vol}(A \setminus R) \leq 1 - (1-h)^r \leq rh$ . Now by the Lipschitz condition (4),

we have  $\sup_{x \in [0,1]^r} |p(x) - \bar{f}_m(x)| \leq Lh\sqrt{r}$  and

$$\begin{aligned}
& |P(A \setminus R) - P_m(A \setminus R)| \\
& \leq \text{number of cubes intersecting } (A \setminus R) \times \text{maximum density discrepancy} \times \text{volume of cube} \\
& \leq (rh/h^r) \cdot (Lh\sqrt{r}) \cdot h^r \leq Lr^{3/2}m^{-2/r}. \tag{21}
\end{aligned}$$

Thus we have by (19), (20) and (21)

$$\mathbb{E} \sup_x |\widehat{F}_m(x) - F(x)| = O\left(\sqrt{\frac{r \log n}{n}}\right) + Lr^{3/2}m^{-2/r}. \tag{22}$$

Hence,

$$\mathbb{E} \sup_x |\widehat{F}_Z(x) - F(x)| = O\left(\sqrt{\frac{r \log k}{k}}\right) + O\left(\sqrt{\frac{r \log n}{n}}\right) + Lr^{3/2}m^{-2/r} + \delta.$$

Set  $m \asymp n^{r/(6+r)}$ ,  $k \asymp m^{4/r} = n^{4/(6+r)}$  and  $\delta = (mk/n\alpha)$  we get for all  $n$  large enough,

$$\mathbb{E} \sup_x |\widehat{F}_Z(x) - F(x)| = O\left(\frac{\sqrt{\log n}}{n^{2/(6+r)}}\right). \quad \square$$

## 9.6 Proof of Theorem 4.3

Let  $\widehat{f}_Z$  be the histogram based on  $Z$  as in (8). Then

$$(\widehat{f}_Z(u) - p(u))^2 \preceq (1 - \delta)^2(p(u) - \widehat{f}_m(u))^2 + \delta^2(p(u) - 1)^2 + (\widehat{f}_{m,\delta}(u) - \widehat{f}_Z)^2$$

where  $\preceq$  means less than, up to constants. Hence,

$$\mathbb{E} \int (\widehat{f}_Z(u) - p(u))^2 du \preceq R_m + \delta^2 + \mathbb{E} \int (\widehat{f}_{m,\delta}(u) - \widehat{f}_Z(u))^2 du$$

where  $R_m$  is the usual  $L_2$  risk of a histogram under the Lipschitz condition (4), namely,  $m^{-2/r} + m/n$ . Conditional on  $X$ ,  $\widehat{f}_Z$  is an unbiased estimate of  $\widehat{f}_m$  with integrated variance  $m/k$ . So,

$$\mathbb{E} \int (\widehat{f}_Z(u) - p(u))^2 du \preceq m^{-2/r} + \frac{m}{n} + \delta^2 + \frac{m}{k}.$$

Minimizing this, subject to (6) yields

$$m \asymp n^{r/(2r+3)}, k \asymp n^{(r+2)/(2r+3)}, \delta \asymp n^{-1/(2r+3)}$$

which yields  $\mathbb{E} \int (\widehat{f}_Z(u) - p(u))^2 du = O(n^{-2/(2r+3)})$ .  $\square$

## 9.7 Proof of Theorem 4.4

(1) Note that  $p - \widehat{f}_Z = p - \widetilde{f} + \widetilde{f} - \widehat{f}_Z = p - \widetilde{f} + O_P\left(\frac{m}{k}\right)$ . When  $k \geq n$ , the latter error is lower order than the other terms and may be ignored. Now,

$$p(x) - \widetilde{f}(x) = p(x) - \widehat{f}_m(x) + \widehat{f}_m(x) - \widetilde{f}(x).$$

Thus

$$\int (p(x) - \widetilde{f}(x))^2 dx \preceq \int (p(x) - \widehat{f}_m(x))^2 dx + \int (\widehat{f}_m(x) - \widetilde{f}(x))^2 dx.$$

The expected value of the first term is the usual risk, namely,  $O(m^{-2/r} + m/n)$ .

For the second term, we proceed as follows. Let  $\widehat{p}_j = C_j/n$  and

$$\widehat{q}_j = \frac{(C_j + \nu_j)_+}{\sum_{s=1}^m (C_s + \nu_s)_+}.$$

We claim that

$$\max_j |\widehat{q}_j - \widehat{p}_j| = O\left(\frac{\log m}{n}\right)$$



almost surely, for all large  $n$ . We have

$$\hat{q}_j = \frac{(C_j + \nu_j)_+}{n} \left( \frac{n}{\sum_{s=1}^m (C_s + \nu_s)_+} \right) = \frac{(C_j + \nu_j)_+}{n} \frac{1}{R_n}$$

where  $R_n = (\sum_{s=1}^m (C_s + \nu_s)_+)/n$ . Now

$$\hat{p}_j - \frac{|\nu_j|}{n} \leq \hat{p}_j + \frac{\nu_j}{n} = \frac{(C_j + \nu_j)}{n} \leq \frac{(C_j + \nu_j)_+}{n} \leq \hat{p}_j + \frac{|\nu_j|}{n}.$$

Therefore,

$$\left| \frac{(C_j + \nu_j)_+}{n} - \hat{p}_j \right| \leq \frac{|\nu_j|}{n} \leq \frac{M}{n}$$

where  $M = \max\{|\nu_1|, \dots, |\nu_m|\}$ . Let  $A > 0$ . The density for  $\nu_j$  has the form  $f(\nu) = (\beta/2)e^{-\beta|\nu|}$ .

So,

$$\mathbb{P}(M > A \log m) \leq m\mathbb{P}(|\nu_j| > A \log m) = \beta m \int_{A \log m}^{\infty} e^{-\beta|\nu|} d\nu = \frac{1}{m^{A\beta-1}}.$$

By choosing  $A$  large enough we have that  $M < A \log m$  a.s. for large  $n$ , by the Borel-Cantelli lemma. Therefore,

$$\left| \frac{(C_j + \nu_j)_+}{n} - \hat{p}_j \right| \leq \frac{\log m}{n}$$

Now we bound  $R_n$ . We have

$$1 - \frac{\sum_s |\nu_s|}{n} \leq 1 + \frac{\sum_s \nu_s}{n} \leq R_n = \frac{\sum_{s=1}^m (C_s + \nu_s)_+}{n} \leq 1 + \frac{\sum_s |\nu_s|}{n}$$

so that

$$|R_n - 1| \leq \frac{\sum_s |\nu_s|}{n} \leq \frac{Mm}{n} = O\left(\frac{m \log m}{n}\right) \quad a.s.$$

Therefore,  $1/R_n = (1 + O(m \log m/n))$  and thus

$$\begin{aligned} \hat{q}_j &= \left( \hat{p}_j + O\left(\frac{\log m}{n}\right) \right) \left( 1 + O\left(\frac{m \log m}{n}\right) \right) \\ &= \hat{p}_j + \hat{p}_j O\left(\frac{m \log m}{n}\right) + O\left(\frac{\log m}{n}\right) + O\left(\frac{m(\log m)^2}{n^2}\right). \end{aligned}$$

Next we claim that  $\hat{p}_j = O(1/m)$  a.s. To see this, note that  $p_j \leq C/m$ , by definition of  $C$ :  $1 \leq C = \sup_x p(x) < \infty$ . Hence, by Bernstein's inequality,

$$\begin{aligned} \mathbb{P}\left(\hat{p}_j > \frac{2C}{m}\right) &= \mathbb{P}\left(\hat{p}_j - p_j > \frac{2C}{m} - p_j\right) \leq \exp\left\{-\frac{1}{2} \frac{n((2C/m) - p_j)^2}{p_j + \frac{1}{3}((2C/m) - p_j)}\right\} \\ &\leq \exp\left\{-\frac{1}{2} \frac{nC^2/m^2}{(4C/3m)}\right\} = e^{-3nC/(8m)} \leq \frac{1}{n^2} \end{aligned}$$

for all  $n \geq 16m \log n/3C$ ; Thus  $\hat{p}_j = O(1/m)$  a.s. for all large  $n$ . Thus,  $\hat{q}_j - \hat{p}_j = O(\log m/n)$  almost surely for all large  $n$ . Hence,

$$\mathbb{E} \int (\hat{f}_m(x) - \tilde{f}(x))^2 dx = O\left(\frac{m \log m}{n}\right)^2.$$

So the risk is

$$O\left(m^{-2/r} + \frac{m}{n} + \left(\frac{m \log m}{n}\right)^2\right) = O\left(m^{-2/r} + \frac{m}{n}\right),$$

for  $n \geq m \log^2 m$ . This is the usual risk. Hence, we can choose  $m \asymp n^{r/(2+r)}$  to achieve risk  $n^{-2/(2+r)}$  for all  $n$  large enough.

(2) Let  $\hat{F}_m$  be the cdf based on the original histogram and let  $\tilde{F}_m$  be the cdf based on the perturbed histogram. We have

$$\begin{aligned} \mathbb{E} \sup_x |F(x) - \hat{F}_Z(x)| &\leq \mathbb{E} \sup_x |F(x) - \hat{F}_m(x)| + \mathbb{E} \sup_x |\hat{F}_m(x) - \tilde{F}_m(x)| + \mathbb{E} \sup_x |\tilde{F}_m(x) - \hat{F}_Z(x)| \\ &\leq \mathbb{E} \sup_x |F(x) - \hat{F}_m(x)| + \mathbb{E} \sup_x |\hat{F}_m(x) - \tilde{F}_m(x)| + O\left(\sqrt{\frac{r \log k}{k}}\right). \end{aligned}$$

Since we may take  $k$  as large as we like, we can make the last term arbitrarily small. From (22),

$$\mathbb{E} \sup_x |F(x) - \hat{F}_m(x)| = O\left(\sqrt{\frac{r \log n}{n}}\right) + Lr^{3/2}m^{-2/r}.$$

Let  $\hat{f}(x) = h^{-r} \sum_{j=1}^m \hat{p}_j I(x \in B_j)$  and Let  $\tilde{f}(x) = h^{-r} \sum_{j=1}^m \hat{q}_j I(x \in B_j)$ . Let  $x' = (u_1 h, \dots, u_r h)$  where  $u_i = \lceil x_i/h \rceil, \forall i = 1, \dots, r$ . Recall that  $B_1, \dots, B_m$  are the  $m$  bins of

$\mathcal{X}$  with sides of length of  $h$ . Let  $B_x$  denote the cube with the left-most corner being 0 and the right-most corner being  $x$ . Then for all  $x$ , we have

$$\begin{aligned} \left| \widehat{F}_m(x) - \widetilde{F}_m(x) \right| &= \left| \int_0^x \widehat{f}(s) - \widetilde{f}(s) ds \right| \leq \int_0^x \left| \widehat{f}(s) - \widetilde{f}(s) \right| ds \\ &\leq \int_0^{x'} \left| \widehat{f}(s) - \widetilde{f}(s) \right| ds \\ &= \sum_{\ell: B_\ell \subseteq B_{x'}} |\widehat{p}_\ell - \widehat{q}_\ell| \leq \sum_{\ell=1}^m |\widehat{p}_\ell - \widehat{q}_\ell| \end{aligned}$$

where we use the fact that there are at most  $m$  cubes. Hence,

$$\mathbb{E} \sup_{x \in [0,1]^r} |\widehat{F}_m(x) - \widetilde{F}_m(x)| \leq \frac{m \log m}{n}$$

where we use the fact that  $\max_j |\widehat{p}_j - \widehat{q}_j| = O(\log m/n)$  a.s. So,

$$\mathbb{E} \sup_x |F(x) - \widehat{F}_Z(x)| = O\left(\sqrt{\frac{r \log n}{n}}\right) + Lr^{3/2}m^{-2/r} + O\left(\frac{m \log m}{n}\right).$$

Setting  $m \asymp n^{r/(2+r)}$  yields

$$\mathbb{E} \sup_x |F(x) - \widehat{F}_Z(x)| = O\left(\min\left(\frac{\log n}{n^{2/(2+r)}}, \sqrt{\frac{\log n}{n}}\right)\right)$$

Hence for  $r = 1$ , the rate is  $O\left(\sqrt{\frac{\log n}{n}}\right)$ . For  $r \geq 2$ , the rate is dominated by the first term inside  $O()$ , and hence the rate is  $O(\log n \times n^{-2/(2+r)})$ .  $\square$

## 9.8 Proof of Theorem 5.3

Let  $B_\epsilon = \left\{ u = (u_1, \dots, u_k) : \rho(F, \widehat{F}_u) \leq \epsilon \right\}$  where  $\widehat{F}_u$  is the empirical distribution based on  $u = (u_1, \dots, u_k) \in \mathcal{X}^k$ . Also, let  $A_n = \{\rho(\widehat{F}_X, F) \leq \epsilon_n/16\}$ . For notational simplicity set

$\Delta = \Delta_{n,k}$ . Then

$$\begin{aligned}
\mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n\right) &= \mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n, A_n\right) + \mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n, A_n^c\right) \\
&\leq \mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n, A_n\right) + \mathbb{P}\left(A_n^c\right) \\
&= \mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n, A_n\right) + O\left(\frac{1}{n^c}\right).
\end{aligned} \tag{23}$$

By the triangle inequality  $\rho(\widehat{F}_u, \widehat{F}_X) \geq \rho(\widehat{F}_u, F) - \rho(\widehat{F}_X, F)$ . Then,

$$\begin{aligned}
\int_{B_\epsilon^c} g_x(u) du &= \int_{B_\epsilon^c} \exp\left(\frac{-\alpha\rho(\widehat{F}_X, \widehat{F}_u)}{2\Delta}\right) du \\
&\leq \int_{B_\epsilon^c} \exp\left(\frac{-\alpha(\rho(\widehat{F}_u, F) - \rho(\widehat{F}_X, F))}{2\Delta}\right) du \\
&= \exp\left(\frac{\alpha\rho(\widehat{F}_X, F)}{2\Delta}\right) \int_{B_\epsilon^c} \exp\left(\frac{-\alpha\rho(\widehat{F}_u, F)}{2\Delta}\right) du \\
&\leq \exp\left(\frac{\alpha\rho(\widehat{F}_X, F)}{2\Delta}\right) \exp\left(\frac{-\alpha\epsilon}{2\Delta}\right) \int_{B_\epsilon^c} du \\
&\leq \exp\left(\frac{\alpha\rho(\widehat{F}_X, F)}{2\Delta}\right) \exp\left(\frac{-\alpha\epsilon}{2\Delta}\right).
\end{aligned}$$

By the triangle inequality, we also have  $\rho(\widehat{F}_u, \widehat{F}_X) \leq \rho(\widehat{F}_u, F) + \rho(\widehat{F}_X, F)$  and

$$\begin{aligned}
\int g_x(u) du &\geq \int_{B_{\epsilon/2}} g_x(u) du = \int_{B_{\epsilon/2}} \exp\left(\frac{-\alpha\rho(\widehat{F}_X, \widehat{F}_u)}{2\Delta}\right) du \\
&\geq \exp\left(\frac{-\alpha\rho(\widehat{F}_X, F)}{2\Delta}\right) \int_{B_{\epsilon/2}} \exp\left(\frac{-\alpha\rho(F, \widehat{F}_u)}{2\Delta}\right) du \\
&\geq \exp\left(\frac{-\alpha\rho(\widehat{F}_X, F)}{2\Delta}\right) \exp\left(\frac{-\alpha\epsilon}{4\Delta}\right) \int_{B_{\epsilon/2}} du \\
&= \exp\left(\frac{-2\alpha\rho(\widehat{F}_X, F) - \alpha\epsilon}{4\Delta}\right) \int_{B_{\epsilon/2}} \frac{p(u_1) \cdots p(u_k)}{p(u_1) \cdots p(u_k)} du \\
&\geq \frac{\exp\left(\frac{-2\alpha\rho(\widehat{F}_X, F) - \alpha\epsilon}{4\Delta}\right)}{(\sup_x p(x))^k} \mathbb{P}\left(\rho(F, \widehat{G}) \leq \epsilon/2\right)
\end{aligned}$$

where  $\widehat{G}$  is the empirical cdf from a sample of size  $k$  drawn from  $P$ . Thus we have

$$\int_{B_\epsilon^c} h(u|x)du \leq \frac{(\sup_x p(x))^k \exp\left(\frac{\alpha\rho(\widehat{F}_X, F)}{\Delta}\right) \exp\left(\frac{-\alpha\epsilon}{4\Delta}\right)}{\mathbb{P}\left(\rho(F, \widehat{G}) \leq \epsilon/2\right)}.$$

Thus, from (23),

$$\begin{aligned} \mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon\right) &\leq \mathbb{P}\left(\rho(\widehat{F}_X, F) \geq \frac{\epsilon}{16}\right) + \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\epsilon}{16\Delta}\right)}{\mathbb{P}\left(\rho(F, \widehat{G}) \leq \epsilon/2\right)} \\ &= \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\epsilon}{16\Delta}\right)}{\mathbb{P}\left(\rho(F, \widehat{G}) \leq \epsilon/2\right)} + O\left(\frac{1}{n^c}\right). \end{aligned}$$

Thus the theorem holds.  $\square$

## 9.9 Proof of Lemma 5.1

**Proof of Lemma 5.1.** We start with KS, By the triangle inequality, we have for all  $z \in \mathcal{X}^k$  and for all  $x, y \in \mathcal{X}^n$ ,

$$\left| \rho(\widehat{F}_x, \widehat{F}_z) - \rho(\widehat{F}_y, \widehat{F}_z) \right| \leq \rho(\widehat{F}_x, \widehat{F}_y).$$

Notice that changing one entry in  $x$  will change  $\widehat{F}_x(t)$  by at most  $\frac{1}{n}$  at any  $t$  by definition, that is,

$$\sup_{t \in [0,1]^r} |\widehat{F}_x(t) - \widehat{F}_y(t)| = \frac{1}{n}.$$

Thus the conclusion holds for the KS-distance.  $\square$

## 9.10 Proof of Theorem 5.4

We need the following small ball result; see Li and Shao (2001).

**Theorem 9.1.** *Let  $r \geq 3$ , and  $\{X_t, t \in [0, 1]^r\}$  be the Brownian sheet. Then there exists  $0 < C_r <$*

$\infty$  such that for all  $0 < \epsilon \leq 1$ ,

$$\log \mathbb{P} \left( \sup_{t \in [0,1]^r} |X_t| \leq \epsilon \right) \geq -C_r \epsilon^{-2} \log^{2r-1}(1/\epsilon)$$

where  $C_r$  depends only on  $r$ . The same bound holds for a Brownian bridge.

*Proof of theorem 5.4.* The Vapnik-Chervonenkis dimension of the class of sets of the form  $\{(-\infty, x_1] \times \cdots \times (-\infty, x_r)\}$  is  $r$  and so by the standard Vapnik-Chervonenkis bound, we have for  $\epsilon_n, k_n$  as specified in the theorem statement,

$$\begin{aligned} \mathbb{P} \left( \sup_{[0,1]^r} |\widehat{F}_X(t) - F(t)| > \frac{\epsilon_n}{16} \right) &\leq 8n^r \exp \left\{ -\frac{n(\epsilon_n/16)^2}{32} \right\} \\ &\leq 8 \exp \left\{ -c_5 \left( \frac{B}{3\alpha} \right)^{2/3} n^{1/3} + r \log n \right\} \\ &= 8 \exp \left\{ -c_6 \sqrt{k_n} \left( \frac{B}{3\alpha} \right) + c_7 r \log k_n \right\} \\ &= 8 \exp \left\{ -C_2 \sqrt{k_n} \left( \frac{B}{3\alpha} \right) \right\} \end{aligned} \quad (24)$$

for some constants  $c_5, c_6, c_7, C_2 > 0$  for  $n$  large enough. Thus (10) holds. Now we compute the small ball probability. Note that  $\sqrt{k}(\widehat{F}_k - F)$  converges to a Brownian bridge  $B_k$  on  $[0, 1]^r$ . More precisely, from Csörgő and Révész (1975) there exist a sequence of Brownian bridges  $B_k$  such that

$$\sup_t |\sqrt{k}(\widehat{F}_k - F)(t) - B_k(t)| = O \left( \frac{(\log k)^{3/2}}{k^\gamma} \right) \quad \text{a.s.} \quad (25)$$

where  $\gamma = 1/(2(r+1))$ . It is clear that the RHS of (25) is  $o(1)$  a.s. given a fixed  $r$ . Hence we have for  $k = k_n$  and  $\epsilon_n$  as chosen in the theorem statement, and for all  $\epsilon \geq \epsilon_n$ , it holds that

$$\begin{aligned} \log \mathbb{P}(\sup_t |\widehat{F}_Z(t) - F(t)| \leq \epsilon/2) &= \log \mathbb{P}(\sup_t \sqrt{k} |\widehat{F}_Z(t) - F(t)| \leq \sqrt{k}\epsilon/2) \\ &\geq \log \mathbb{P} \left( \sup_t |B_k(t)| \leq \sqrt{k}\epsilon - O(k^{-\gamma}(\log k)^{3/2}) \right) \end{aligned} \quad (26)$$

$$\geq \log \mathbb{P} \left( \sup_t |B_k(t)| \leq \frac{\sqrt{k}\epsilon}{4} \right) \quad (27)$$

for all large  $n$ , where (26) follows from (25) and (27) holds given that  $\sqrt{k}\epsilon \geq \sqrt{k_n}\epsilon_n \geq c$  for some constant  $c > 1/2$  due to our choice of  $k_n$  and  $\epsilon_n$ . Also,  $\Delta \leq 1/n$  for KS distance. Hence, by Theorem 5.3 and (24), we have for  $B = \log \sup_x p(x) > 0$ ,

$$\begin{aligned} & \mathbb{P}\left(\rho(F, \widehat{F}_Z) > \epsilon_n\right) \\ & \leq C_0 \exp\left\{-n\left(\frac{3\alpha\epsilon_n}{16} - \frac{Bk_n}{n} - \frac{C_1|\log(\sqrt{k_n}\epsilon_n/4)|^{2r-1}}{nk_n\epsilon_n^2}\right)\right\} + 8 \exp\left\{-C_2\frac{B\sqrt{k_n}}{3\alpha}\right\} \\ & \leq C_0 \exp(-C_3Bk_n/2) + 8 \exp\left\{-C_2\left(\frac{B}{3\alpha}\right)\sqrt{k_n}\right\} \rightarrow 0 \end{aligned} \quad (28)$$

for some constants  $C_0, C_1, C_2$  and  $C_3$ , where (28) holds when we take w.l.o.g.  $k_n = \frac{1}{16}\left(\frac{3\alpha}{B}\right)^{2/3}n^{2/3}$  and  $\epsilon_n \geq 2\left(\frac{B}{3\alpha}\right)^{1/3}n^{-1/3}$ , given that  $\epsilon_n \geq 2\left(\frac{B}{3\alpha}\right)^{1/3}n^{-1/3} = \frac{32k_nB}{3n\alpha}$  and hence  $\frac{3\alpha\epsilon_n}{16} \geq \frac{2Bk_n}{n}$ . Thus the result follows.  $\square$

**Remark 9.2.** *The constants taken in the proof are arbitrary; indeed, when we take  $k_n = C_4\left(\frac{3\alpha}{B}\right)^{2/3}n^{2/3}$  and  $\epsilon_n = 32C_4\left(\frac{B}{3\alpha}\right)^{1/3}n^{-1/3}$  with some constant  $C_4 \geq 1/16$ , (28) will hold with slightly different constants  $C_2, C_3$ . For  $k_n$  and  $\epsilon_n$  as chosen above, it holds that  $\sqrt{k_n}\epsilon_n \asymp 1$ .*

## 9.11 Proofs for Lemma 6.1 and Theorem 6.2

Throughout this section, we let  $\widehat{p}_X$  denote the estimator as defined in (14), which is based on a sample of size  $n$  drawn independently from  $F$ ; Similarly, we let  $\widehat{p}_k$  denote the same estimator based on an i.i.d. sample  $(Y_1, \dots, Y_k)$  of size  $k$  drawn from  $F$ , with  $m_k = k^{1/(2\gamma+1)}$  replacing  $m_n$  and  $\widehat{\beta}_j = k^{-1}\sum_{i=1}^k\psi_j(Y_i)$  in (14). We let  $\widehat{p}_Z$  denote the estimator as in (16), based on an i.i.d. sample  $Z = (Z_1, \dots, Z_k)$  of size  $k$  drawn from  $g_x(z)$  as in (17).

**Proof of Lemma 6.1.** Without loss of generality, let  $X = (x, X_2, \dots, X_n)$  and  $Y = (y, X_2, \dots, X_n)$  so that  $\delta(X, Y) = 1$  and let  $Z \in \mathcal{X}^k$ . Recall that

$$\begin{aligned} \xi(X, Z) &= \left(\int(\widehat{p}_X(x) - \widehat{p}_Z(x))^2 dx\right)^{1/2}, \\ \xi(Y, Z) &= \left(\int(\widehat{p}_Y(x) - \widehat{p}_Z(x))^2 dx\right)^{1/2}. \end{aligned}$$

In particular, let us define  $u = \widehat{p}_X - \widehat{p}_Z$  and  $v = \widehat{p}_Y - \widehat{p}_Z$  and thus

$$\begin{aligned}
|\xi(X, Z) - \xi(Y, Z)| &= \left| \left( \int (\widehat{p}_X(x) - \widehat{p}_Z(x))^2 dx \right)^{1/2} - \left( \int (\widehat{p}_Y(x) - \widehat{p}_Z(x))^2 dx \right)^{1/2} \right| \\
&= \left| \|u\|_{\ell_2} - \|v\|_{\ell_2} \right| \leq \|u - v\|_{\ell_2} \\
&= \|\widehat{p}_X - \widehat{p}_Z - (\widehat{p}_Y - \widehat{p}_Z)\|_{\ell_2} = \|\widehat{p}_X - \widehat{p}_Y\|_{\ell_2} \leq \frac{2c_0^2 m_n}{n},
\end{aligned}$$

where the first inequality is due to the triangle inequality for the  $\|\cdot\|_{\ell_2}$  and the last step is due to

$$\begin{aligned}
|\widehat{p}_X(x) - \widehat{p}_Y(x)| &= \frac{1}{n} \left| \sum_{j=1}^{m_n} \left( \sum_{i=1}^n \psi_j(X_i) - \sum_{i=1}^n \psi_j(Y_i) \right) \psi_j(x) \right| \\
&= \frac{1}{n} \left| \sum_{j=1}^{m_n} (\psi_j(X_1) - \psi_j(Y_1)) \psi_j(x) \right| \\
&\leq \frac{1}{n} \sum_{j=1}^{m_n} (|\psi_j(X_1)| + |\psi_j(Y_1)|) |\psi_j(x)| \leq \frac{2c_0^2 m_n}{n}.
\end{aligned}$$

Hence  $\Delta \leq \frac{2c_0^2 m_n}{n}$ .  $\square$

**Proof of Theorem 6.2.** For  $u = (u_1, \dots, u_k) \in \mathcal{X}^k$ , we let

$$\widehat{p}_u(x) = 1 + \sum_{j=1}^{m_k} \widehat{\beta}_j \psi_j(x),$$

where  $m_k = k^{\frac{1}{2\gamma+1}}$  and  $\widehat{\beta}_j = k^{-1} \sum_{i=1}^k \psi_j(u_i)$ .

Let  $\widehat{F}_u$  be the empirical distribution based on  $u$ . Our proof follows that of Theorem 5.3, with

$$\rho(F, \widehat{F}_u) = \|p - \widehat{p}_u\|_{\ell_2} \quad \text{and} \quad \rho(F_X, \widehat{F}_u) = \|\widehat{p}_X - \widehat{p}_u\|_{\ell_2}$$

as defined in (15) for  $X = (X_1, \dots, X_n)$ . Now

$$B_\epsilon = \left\{ u = (u_1, \dots, u_k) : \|p - \widehat{p}_u\|_{\ell_2} < \epsilon \right\}.$$



Thus the corresponding triangle inequalities that we use to replace that in Theorem 5.3 are:

$$\begin{aligned}\|\widehat{p}_u - \widehat{p}_X\|_{\ell_2} &\geq \|\widehat{p}_u - p\|_{\ell_2} - \|\widehat{p}_X - p\|_{\ell_2} \quad \text{and} \\ \|\widehat{p}_u - \widehat{p}_X\|_{\ell_2} &\leq \|\widehat{p}_u - p\|_{\ell_2} + \|p - \widehat{p}_X\|_{\ell_2}.\end{aligned}$$

Standard risk calculations show that (10) holds for some  $c > 0$  with  $\rho(F, \widehat{F}_X)$  being replaced with  $\|\widehat{p}_X - p\|_{\ell_2}$ . That is, by Markov's inequality,

$$\mathbb{P}(\|\widehat{p}_X - p\|_{\ell_2} > \epsilon) \leq \frac{\mathbb{E} \|\widehat{p}_X - p\|_{\ell_2}^2}{\epsilon^2}$$

and (10) follows from the polynomial decay of the mean squared error  $\mathbb{E}\|\widehat{p}_X - p\|^2$ . Thus, from (23), for  $\widehat{p}_Z = \widehat{p}^*$  as in (16),

$$\begin{aligned}\mathbb{P}(\|p - \widehat{p}_Z\|_{\ell_2} > \epsilon) &\leq \mathbb{P}\left(\|\widehat{p}_X - p\|_{\ell_2} \geq \frac{\epsilon}{16}\right) + \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\epsilon}{16\Delta}\right)}{\mathbb{P}(\|p - \widehat{p}_k\|_{\ell_2} \leq \epsilon/2)} \\ &= \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\epsilon}{16\Delta}\right)}{\mathbb{P}(\|p - \widehat{p}_k\|_{\ell_2} \leq \epsilon/2)} + O\left(\frac{1}{n^c}\right).\end{aligned}$$

We need to compute the small ball probability. Recall that  $\widehat{p}_k$  denote the estimator based on a sample of size  $k$ . By Parseval's relation,

$$\int (p(x) - \widehat{p}_k(x))^2 dx = \sum_{j=1}^{m_k} (\widehat{\beta}_j - \beta_j)^2 + \sum_{m_k+1}^{\infty} \beta_j^2 \leq \sum_{j=1}^{m_k} (\widehat{\beta}_j - \beta_j)^2 + ck^{-2\gamma/(2\gamma+1)}.$$

Let  $U_i = (\psi_1(X_i) - \beta_1, \dots, \psi_{m_k}(X_i) - \beta_{m_k})^T$  and  $Y_i = \Sigma_k^{-1/2} U_i$  where  $\Sigma_k$  is the covariance matrix of  $U_i$ . Hence,  $Y_i$  has mean 0 and identity covariance matrix. Let  $\lambda_k$  denote the largest eigenvalue of  $\Sigma_k$ . From Lemma 9.3 below,  $\lambda = \limsup_{k \rightarrow \infty} \lambda_k < \infty$ . Let  $Q = \sum_{j=1}^{m_k} (\widehat{\beta}_j - \beta_j)^2$  and let  $S = k^{-1/2} \sum_{i=1}^k Y_i$ . Then, for all large  $k$ , and any  $\delta > 0$ ,

$$\mathbb{P}(Q \leq \delta^2) = \mathbb{P}(S^T \Sigma_k S \leq k\delta^2) \geq \mathbb{P}\left(S^T S \leq \frac{k\delta^2}{\lambda_k}\right) \geq \mathbb{P}\left(S^T S \leq \frac{k\delta^2}{2\lambda}\right).$$

From Theorem 1.1 of Bentkus (2003) we have that

$$\sup_c |\mathbb{P}(S^T S \leq c) - \mathbb{P}(\chi_{m_k}^2 \leq c)| = O\left(\sqrt{\frac{m_k^3}{k}}\right) = O(k^{-(\gamma-1)/(2\gamma+1)}).$$

Next we use the fact (see Rohde and Duembgen (2008) for example) that  $\mathbb{P}(\chi_m^2 \leq m+a) \geq 1 - e^{-a^2/(4(m+a))}$ . Let  $k = \sqrt{n}$ ,  $\epsilon_n = c_1 n^{-\gamma/(2\gamma+1)}$  where  $c_1 \geq 4(2\lambda + 1)(C^2 + 1)$

$$a = \frac{k(\epsilon_n/4 - C^2 k^{-2\gamma/(2\gamma+1)})}{2\lambda} - m_k \geq (C^2 + 1)n^{1/2(2\gamma+1)} - m_k \geq C^2 m_k,$$

since  $m_k = k^{\frac{1}{2\gamma+1}} = n^{1/2(2\gamma+1)}$ . We see that for all large  $k$

$$\begin{aligned} \mathbb{P}\left(\|p - \hat{p}_k\|_{\ell_2} \leq \frac{\sqrt{\epsilon_n}}{2}\right) &= \mathbb{P}\left(\int (p(x) - \hat{p}_k(x))^2 dx \leq \frac{\epsilon_n}{4}\right) \\ &\geq \mathbb{P}\left(\sum_{j=1}^{m_k} (\hat{\beta}_j - \beta_j)^2 \leq \frac{\epsilon_n}{4} - C^2 k^{-2\gamma/(2\gamma+1)}\right) \\ &= \mathbb{P}\left(\chi_{m_k}^2 \leq \frac{k(\epsilon_n/4 - C^2 k^{-2\gamma/(2\gamma+1)})}{2\lambda}\right) - O(k^{-(\gamma-1)/(2\gamma+1)}) \\ &\geq 1 - \exp\left(\frac{-a^2}{4(m_k + a)}\right) - O(k^{-(\gamma-1)/(2\gamma+1)}) \\ &\geq \frac{1}{2} - O(k^{-(\gamma-1)/(2\gamma+1)}). \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{P}(\|p - \hat{p}_Z\|_{\ell_2} > \sqrt{\epsilon_n}) &\leq \mathbb{P}\left(\|\hat{p}_X - p\|_{\ell_2} \geq \frac{\sqrt{\epsilon_n}}{16}\right) + \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\sqrt{\epsilon_n}}{16\Delta}\right)}{\mathbb{P}(\|p - \hat{p}_k\|_{\ell_2} \leq \sqrt{\epsilon_n}/2)} \\ &= \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha\sqrt{\epsilon_n}}{16\Delta}\right)}{\mathbb{P}(\|p - \hat{p}_k\|_{\ell_2} \leq \sqrt{\epsilon_n}/2)} + O\left(\frac{1}{n^c}\right) \\ &\leq \frac{(\sup_x p(x))^k \exp\left(\frac{-3\alpha n \sqrt{\epsilon_n}}{32c_0^2 m_n}\right)}{\mathbb{P}(\|p - \hat{p}_k\|_{\ell_2} \leq \sqrt{\epsilon_n}/2)} + O\left(\frac{1}{n^c}\right) \end{aligned}$$

and so for  $\gamma > 1$ ,

$$\begin{aligned}
\mathbb{P}\left(\int (\widehat{p}_Z - p)^2 \leq \epsilon_n\right) &\leq c_2 \exp\left(k \log \sup_x p(x)\right) \exp\left(\frac{-3\sqrt{c_1}\alpha n}{n^{1/(2\gamma+1)}n^{\gamma/2(2\gamma+1)}}\right) \\
&= c_2 \exp\left(n^{1/2} \log \sup_x p(x) - \alpha c_3 n^{\left(\frac{3\gamma}{2(2\gamma+1)}\right)}\right) \\
&= c_2 \exp\left(-\alpha c_4 n^{\left(\frac{3\gamma}{2(2\gamma+1)}\right)}\right) \rightarrow 0,
\end{aligned}$$

as  $n \rightarrow \infty$  since  $\frac{3\gamma}{2(2\gamma+1)} > 1/2$ , where  $c_2, c_3, c_4$  are some constants. Hence the theorem holds.  $\square$

**Lemma 9.3.** *Let  $\lambda = \limsup_{k \rightarrow \infty} \lambda_k$ . Then  $\lambda < \infty$ .*

*Proof.* Recall that the orthonormal basis is  $\psi_0, \psi_1, \dots$ , where  $\psi_0 = 1$  and  $\psi_j(x) = \sqrt{2} \cos(\pi j x)$ . Also  $p(x) = 1 + \sum_{j=1}^{\infty} \beta_j \psi_j(x)$  and  $\sum_j \beta_j^2 j^{2\gamma} < \infty$ . Note that  $\sum_{j=1}^{\infty} |\beta_j|^k = O(1)$  for  $k \geq 1$ ; see Efromovich (1999). Note that  $\Sigma_k$  is the covariance matrix of  $\widehat{\beta}$  times  $n$ . We will use the standard identities  $\cos^2(u) = (1 + \cos(2u))/2$  and  $\cos(u) \cos(v) = \frac{\cos(u-v) + \cos(u+v)}{2}$ . It follows that  $\psi_j^2(x) = 1 + \frac{1}{\sqrt{2}} \psi_{2j}(x)$  and  $\psi_j(x) \psi_k(x) = \frac{\psi_{j-k}(x) + \psi_{j+k}(x)}{\sqrt{2}}$ . Now  $\mathbb{E}(\widehat{\beta}_j) = \beta_j$ . And

$$n \text{Var}(\widehat{\beta}_j) = \text{Var}(\psi_j(X)) = \int \psi_j^2(x) p(x) dx - \beta_j^2.$$

Now  $\int \psi_j^2(x) p(x) dx = \int \psi_j^2(x) (1 + \sum_{\ell=1}^{\infty} \beta_{\ell} \psi_{\ell}(x)) dx = 1 + \sum_{\ell=1}^{\infty} \beta_{\ell} \int \psi_{\ell}(x) \psi_j^2(x) dx = 1 + \frac{1}{2} \sum_{\ell=1}^{\infty} \beta_{\ell} \int \psi_{\ell}(x) \left(1 + \frac{\psi_{2j}(x)}{\sqrt{2}}\right) dx = 1 + \frac{\beta_{2j}}{\sqrt{2}}$ . Thus,  $\Sigma_{jj} = 1 + \frac{\beta_{2j}}{\sqrt{2}} - \beta_j^2$ . Now consider  $j \neq k$ .

Then

$$\begin{aligned}
\mathbb{E}(\psi_j(X)\psi_k(X)) &= \int \psi_j(x)\psi_k(x)p(x)dx \\
&= \sum_{\ell} \beta_{\ell} \int \psi_j(x)\psi_k(x)dx \\
&= \beta_j \int \psi_j^2(x)\psi_k(x)dx + \beta_k \int \psi_k^2(x)\psi_j(x)dx + \sum_{\ell \neq j,k} \beta_{\ell} \int \psi_j(x)\psi_k(x)\psi_{\ell}(x)dx \\
&= \frac{\beta_j}{\sqrt{2}} \int \psi_{2j}(x)\psi_k(x)dx + \frac{\beta_k}{\sqrt{2}} \int \psi_{2k}(x)\psi_j(x)dx \\
&\quad + \frac{1}{\sqrt{2}} \sum_{\ell \neq j,k} \beta_{\ell} \int (\psi_{j-k}(x) + \psi_{j+k}(x))\psi_{\ell}(x) \\
&= \frac{\beta_j}{\sqrt{2}} I(2j = k) + \frac{\beta_k}{\sqrt{2}} I(2k = j) \\
&\quad + \frac{\beta_{\ell}}{\sqrt{2}} I(\ell = |j - k| \ \& \ j \neq 2k) + \frac{\beta_{\ell}}{\sqrt{2}} I(\ell = j + k) \\
&= \frac{\beta_k}{\sqrt{2}} I(2k = j) + \frac{\beta_{|j-k|}}{\sqrt{2}} I(j \neq 2k) + \frac{\beta_{j+k}}{\sqrt{2}} \\
&= \frac{\beta_{|j-k|}}{\sqrt{2}} + \frac{\beta_{j+k}}{\sqrt{2}},
\end{aligned}$$

where we used the fact that  $\psi_{-j}(x) = \psi_j(x)$  for all  $j = 1, 2, \dots$  and  $\int \psi_j(x)dx = 0$  for all  $j > 0$ .

So, we have for all  $j \in \{1, \dots, p\}$ ,

$$\begin{aligned}
\sum_{k=1}^p |\Sigma_{jk}| &= |\Sigma_{jj}| + \sum_{j \neq k} \left| \frac{\beta_{|j-k|}}{\sqrt{2}} + \frac{\beta_{j+k}}{\sqrt{2}} - \beta_j \beta_k \right| \\
&\leq 1 + \left| \frac{\beta_{2j}}{\sqrt{2}} \right| + |\beta_j| \sum_k |\beta_k| + \sum_{j \neq k} \left| \frac{\beta_{|j-k|}}{\sqrt{2}} \right| + \left| \frac{\beta_{j+k}}{\sqrt{2}} \right| \\
&\leq 1 + \left| \frac{\beta_{2j}}{\sqrt{2}} \right| + (|\beta_j| + \sqrt{2}) \sum_{k=1}^{\infty} |\beta_k| \\
&= O(1).
\end{aligned}$$

Hence,  $\limsup_{k \rightarrow \infty} \lambda_{\max}(\Sigma_k) \leq \|\Sigma_k\|_{\infty} = O(1)$  and the lemma holds.  $\square$

## 9.12 Proof of Theorem 6.3

The proof is similar to the proof of Theorem 4.4, so we provide a short outline. In particular, the effect of truncation can be shown to be negligible as in the proof of Theorem 4.4. We have  $p - \hat{p}_Z = p - \hat{q} + \hat{q} - \hat{p}_Z = p - \hat{q} + O_P(m/k)$  and the latter term is negligible for  $k \geq n$ . Now  $p - \hat{q} = p - \hat{p} + \hat{p} - \hat{q}$ . The term  $p - \hat{p}$  is the usual error term and contributes  $O(n^{-2\gamma/(2\gamma+1)})$  to the risk. For the second term,  $\int(\hat{p} - \hat{q})^2 = \sum_{j=1}^m \nu_j^2 = O_P(m/n) = O_P(n^{-2\gamma/(2\gamma+1)})$ .  $\square$

## References

- AGGARWAL, G., FEDER, T., KENTHAPADI, K., KHULLER, S., PANIGRAHY, R., THOMAS, D. and ZHU, A. (2006). Achieving anonymity via clustering. *Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 153–162.
- BARAK, B., CHAUDHURI, K., DWORK, C., KALE, S., MCSHERRY, F. and TALWAR, K. (2007). Privacy, accuracy, and consistency too: a holistic solution to contingency table release. *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 273–282.
- BENTKUS, V. (2003). On the dependence of the berryseen bound on dimension. *Journal of Statistical Planning and Inference* **113** 385–402.
- BLUM, A., DWORK, C., MCSHERRY, F. and NISSIM, K. (2005). Practical privacy: the SuLQ framework. *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 128–138.
- BLUM, A., LIGETT, K. and ROTH, A. (2008). A Learning Theory Approach to Non-Interactive Database Privacy. *Proceedings of the 40th annual ACM symposium on Theory of computing* 609–618.
- CSÖRGŐ, M. and RÉVÉSZ, P. (1975). A new method to prove strassen type laws of invariance principle. II. *Probability Theory and Related Fields* 261–269.

- DINUR, I. and NISSIM, K. (2003). Revealing information while preserving privacy. *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* 202–210.
- DUNCAN, G. and LAMBERT, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association* 10–28.
- DUNCAN, G. and LAMBERT, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* 207–217.
- DUNCAN, G. and PEARSON, R. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science* **6** 219–232.
- DWORK, C. (2006). Differential privacy. *33rd International Colloquium on Automata, Languages and Programming* 1–12.
- DWORK, C. and LEI, J. (2009). Differential privacy and robust statistics. *Proceedings of the 41st ACM Symposium on Theory of Computing* 371–380.
- DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. *Proceedings of the 3rd Theory of Cryptography Conference* 265–284.
- DWORK, C., MCSHERRY, F. and TALWAR, K. (2007). The price of privacy and the limits of LP decoding. *Proceedings of the 39th annual ACM symposium on Theory of computing* 85–94.
- DWORK, C., NAOR, M., REINGOLD, O., ROTHBLUM, G. and VADHAN, S. (2009). On the complexity of differentially private data release. *Proceedings of the 41st ACM Symposium on Theory of Computing* 381–390.
- DWORK, C. and NISSIM, K. (2004). Privacy-preserving datamining on vertically partitioned databases. *Proceedings of the 24th Annual International Cryptology Conference –CRYPTO* 528–544.
- EFROMOVICH, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer-Verlag.

- EVFIMIEVSKI, A., SRIKANT, R., AGRAWAL, R. and GEHRKE, J. (2004). Privacy preserving mining of association rules. *Information Systems* **29** 343 – 364.
- FEIGENBAUM, J., ISHAI, Y., MALKIN, T., NISSIM, K., STRAUSS, M. J. and WRIGHT, R. N. (2006). Secure multiparty computation of approximations. *ACM Trans. Algorithms* **2** 435–472.
- FELDMAN, D., FIAT, A., KAPLAN, H. and NISSIM, K. (2009). Private coresets. *Proceedings of the 41st ACM Symposium on Theory of Computing* 361–370.
- FIENBERG, S. and MCINTYRE, J. (2004). Data Swapping: Variations on a Theme by Dalenius and Reiss. *Privacy in Statistical Databases* **3050** 14–29.
- FIENBERG, S. E., KARR, A. F., NARDI, Y. and SLAVKOVIC, A. (2007). Secure logistic regression with distributed databases. *Bulletin of the ISI* .
- FIENBERG, S. E., MAKOV, U. E. and STEELE, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data (with discussion). *Journal of Official Statistics* **14** 485–511.
- GANTA, S., KASIVISWANATHAN, S. and SMITH, A. (2008). Composition attacks and auxiliary information in data privacy. *Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 265–273 CoRR abs/0803.0032: (2008).
- GHOSH, A., ROUGHGARDEN, T. and SUNDARARAJAN, M. (2009). Universally utility-maximizing privacy mechanisms. *Proceedings of the 41st ACM Symposium on Theory of Computing* 351–360.
- HALL, P. and MURISON, R. D. (1993). Correcting the negativity of high-order kernel density estimators. *Journal of Multivariate Analysis* **47** 103–122.
- KASIVISWANATHAN, S., LEE, H., NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2008). What Can We Learn Privately? *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science* 531–540.

- KIM, J. J. and WINKLER, W. E. (2003). Multiplicative noise for masking continuous data. Tech. rep., Statistical Research Division, US Bureau of the Census, Washington D.C.
- LI, N., LI, T. and VENKATASUBRAMANIAN, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *Proceedings of the 23rd International Conference on Data Engineering* 106–115.
- LI, W. and SHAO, Q.-M. (2001). Gaussian processes: Inequalities, small ball probabilities and applications. In *STOCHASTIC PROCESSES: THEORY AND METHODS. Handbook of Statistics* (C. Rao and D. Shanbhag, eds.), vol. 19. Elsevier, 533–598.
- MACHANAVAJJHALA, A., GEHRKE, J., KIFER, D. and VENKITASUBRAMANIAM, M. (2006).  $\ell$ -diversity: Privacy beyond kappa-anonymity. *Proceedings of the 22nd International Conference on Data Engineering* 24.
- MACHANAVAJJHALA, A., KIFER, D., ABOWD, J., GEHRKE, J. and VILHUBER, L. (2008). Privacy: Theory meets Practice on the Map. *Proceedings of the 24th International Conference on Data Engineering* 277–286.
- MCSHERRY, F. and TALWAR, K. (2007). Mechanism Design via Differential Privacy. *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science* 94–103.
- NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2007). Smooth sensitivity and sampling in private data analysis. *Proceedings of the 39th annual ACM annual ACM symposium on Theory of computing* 75–84.
- PINKAS, B. (2002). Cryptographic techniques for privacy-preserving data mining. *ACM SIGKDD Explorations Newsletter* 4.
- RASTOGI, V., HAY, M., MIKLAU, G. and SUCIU, D. (2009). Relationship privacy: Output perturbation for queries with joins. *Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2009* 107–116.



- REITER, J. (2005). Estimating risks of identification disclosure for microdata. *Journal of the American Statistical Association* **100** 1103 – 1113.
- ROHDE, A. and DUENBGEN, L. (2008). Confidence sets for the optimal approximating model - bridging a gap between adaptive point estimation and confidence regions. *arXiv:0802.3276v2 [math.ST]*.
- SANIL, A. P., KARR, A., LIN, X. and REITER, J. P. (2004). Privacy preserving regression modelling via distributed computation. *Proceedings of Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 677–682.
- SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.
- SMITH, A. (2008). Efficient, differentially private point estimators. ArXiv:0809.4794v1.
- SWEENEY, L. (2002). k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10** 557–579.
- TING, D., FIENBERG, S. E. and TROTTINI, M. (2008). Random orthogonal matrix masking methodology for microdata release. *Int. J. of Information and Computer Security* **2** 86–105.
- WARNER, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60** 63–69.

**ARIZONA INDEPENDENT REDISTRICTING COMMISSION**

*Assorted Materials Supporting the U.S. Census Bureau's Use  
of Differential Privacy*

## Table of Contents

Defs.’ Response in Opposition, <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.).....	653
Dec. of John M. Abowd, <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.).....	737
Amicus Brief of Data Privacy Experts, <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.) .....	859
Supplemental Dec. of John M. Abowd, <i>Alabama v. U.S. Dep’t. of Commerce</i> , No. 3:21-CV-211 (M.D. Ala.) .....	888
John M. Abowd et al., <i>The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau</i> (Aug. 2020) .....	911
Aloni Cohen, et al., <i>Census TopDown: The Impacts of Differential Privacy on Redistricting</i> .....	927
<i>Formal Privacy Methods for the 2020 Census</i> , JASON (Mar. 29, 2020) .....	949
Sam Wang et al., <i>Comment on “The Impact of the U.S. Census Disclosure Avoidance System on Redistricting and Voting Rights Analysis” by Kenny et al.</i> , ELECTORAL INNOVATION LAB, PRINCETON UNIV. (June 2, 2021).....	1099
Michael Hawes, <i>Understanding the 2020 Census Disclosure Avoidance System: Differential Privacy 101</i> , U.S CENSUS BUREAU (May 4, 2021) .....	1105
Michael Hawes et al., <i>Understanding the 2020 Census Disclosure Avoidance System: Differential Privacy 201 and the TopDown Algorithm</i> , U.S CENSUS BUREAU (May 13, 2021) .....	1136
Tommy Wright et al., <i>Empirical Study of Two Aspects of the Topdown Algorithm Output for Redistricting: Reliability &amp; Variability</i> , U.S CENSUS BUREAU (May 18, 2021) .....	1160
Michael Hawes et al., <i>Determining the Privacy-loss Budget</i> , U.S CENSUS BUREAU (June 4, 2021) .....	1322
Release, <i>Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results</i> , U.S CENSUS BUREAU (May 13, 2021) .....	1353

**UNITED STATES DISTRICT COURT  
FOR THE MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION**

THE STATE OF ALABAMA, *et al.*

Plaintiffs,

v.

UNITED STATES DEPARTMENT OF  
COMMERCE, *et al.*,

Defendants.

No. 3:21-cv-00211-RAH-ECM-KCN

**DEFENDANTS' RESPONSE IN OPPOSITION TO  
PLAINTIFFS' MOTION FOR PRELIMINARY INJUNCTION  
AND PETITION FOR WRIT OF MANDAMUS**

## TABLE OF CONTENTS

INTRODUCTION .....	1
BACKGROUND.....	4
A.    The Decennial Census .....	4
B.    The Census Act’s Confidentiality Provisions .....	5
C.    The Rise of Computing Power and Its Implications for Confidentiality .....	6
D.    Differential Privacy.....	8
E.    The Census Bureau’s Delivery of Redistricting Data .....	16
ARGUMENT.....	19
I.    Plaintiffs Lack Standing .....	19
A.    Plaintiffs Have Not Sustained Any Injuries-in-Fact .....	20
1.    Plaintiffs Are Not Injured by Differential Privacy .....	20
a.    Informational Injury .....	20
b.    Sovereign Injury .....	29
c.    Federal Funding .....	33
d.    Vote Dilution .....	35
e.    Section 209.....	36
2.    Plaintiffs Are Not Injured by Delayed Redistricting Data .....	36
B.    Plaintiffs’ Alleged Injuries Are Not Traceable to Defendants’ Actions .....	38
1.    Plaintiffs’ Alleged Injuries Cannot Be Traced to Defendants’ Plan to Use Differential Privacy .....	38
2.    Plaintiffs’ Alleged Injuries Cannot Be Traced to Defendants’ Delay in Producing Redistricting Data .....	40

C.	Plaintiffs’ Purported Injuries Are Not Redressable .....	40
1.	Enjoining Differential Privacy Would Not Redress Plaintiffs’ Alleged Injuries .....	41
2.	Requiring the Census Bureau to Produce Redistricting Data Sooner Would Not Redress Plaintiffs’ Alleged Injuries .....	42
II.	Plaintiffs Are Not Entitled to a Preliminary Injunction.....	44
A.	Plaintiffs Are Unlikely to Succeed on the Merits of Their Differential Privacy Claims.....	45
1.	Plaintiffs’ Census Act Claim Is Not Likely to Succeed .....	45
2.	The Individual Plaintiffs’ Equal Protection Claim Is Not Likely to Succeed .....	48
3.	Plaintiffs’ APA Challenges to Differential Privacy Are Not Likely To Succeed.....	49
a.	The Differential Privacy Announcement Was Not Final Agency Action .....	49
b.	Even Assuming the Differential Privacy Announcement Constituted Final Agency Action, It Did Not Violate the APA .....	55
4.	The Doctrine of Laches Bars Plaintiffs’ Differential Privacy Claims .....	59
B.	Plaintiffs’ Challenge to the February 12 Press Release Is Not Likely to Succeed. ....	61
1.	Plaintiffs’ Claim that the Press Release “Violates the Census Act” Is Not Likely to Succeed.....	61
2.	Alabama’s APA Challenge to the February 12 Press Release Is Not Likely to Succeed.....	64
a.	The February 12 Press Release Was Not Final Agency Action.....	64
b.	The February 12 Press Release is Not Arbitrary or Capricious .....	66

C.	Plaintiffs Will Suffer No Harm, Much Less Irreparable Harm. ....	68
1.	Plaintiffs Have Not Established Irreparable Harm Due to Differential Privacy .....	68
2.	Plaintiffs Have Not Established Irreparable Harm on Their Delay Claim.....	71
D.	Defendants and the Public Would Be Harmed by an Injunction. ....	73
III.	Mandamus Relief Is Unavailable. ....	75
	CONCLUSION .....	78

## INTRODUCTION

Every decade, the United States Census Bureau has the responsibility of “counting the whole number of persons in each State.” U.S. Const. amend. XIV, § 2. Counting over 330 million people across 3.8 million square miles is a very difficult and complex task. Each decennial census takes over a decade to plan, execute, and complete, and involves myriad operational decisions. The 2020 decennial census—a 15.6-billion-dollar operation—is monitored and managed using a master schedule with over 27,000 separate lines of census activities, and is supported by no fewer than 52 separate information-technology systems.

The decennial census is also very important. It underpins our Nation’s representative democracy. It is used to allocate political power at all levels of government. And the data it collects and produces are used for countless purposes by governments, businesses, organizations, and individuals. Given the importance of the census, the Census Bureau must proceed carefully, with meticulous planning. Systems are developed, and tested, and tested again.

None of this would be possible without the cooperation of the public at large. Members of the public can be reluctant to reveal their and their household’s personal information to the government. But we ask them to do so every decade based on the promise—printed at the top of the census questionnaire—that their responses “are protected by law.”

This lawsuit concerns two large obstacles to the successful operation of the 2020 decennial census. The first obstacle is the COVID-19 pandemic, which unfortunately emerged just as hundreds of thousands of census field staff prepared to fan out around the country to collect information from the public. The once-in-a-century pandemic, along with major hurricanes and wildfires, caused a series of cascading delays that has rendered the Census Bureau unable to meet the statutory deadlines for delivering apportionment and redistricting data.



The second obstacle is the rise of computational power that threatens to reveal confidential information. It is now possible, using sophisticated algorithms on powerful systems, to reverse-engineer large sets of aggregated, supposedly de-identified data. Given this development, the Census Bureau set out to determine whether its data products were susceptible to such a “reconstruction attack.” And the Census Bureau determined—and third parties have confirmed—that the disclosure-avoidance method the Bureau applied to protect its 2010 data products no longer suffices to protect the confidentiality of census responses. If the Census Bureau were to continue doing what it did in 2010, it would be violating not only federal law, but also the confidentiality promise that it made to census respondents. And with that bond of trust broken, future census response rates would undoubtedly fall, and the accuracy of future censuses would suffer.

Plaintiffs—the State of Alabama, a congressional representative, and two individuals—would impose a third obstacle to the Census Bureau’s operations if the relief they seek through this lawsuit were granted. Plaintiffs first argue that the disclosure-avoidance method that the Census Bureau will apply to its forthcoming redistricting data products—differential privacy—will result in flawed numbers. They attempt to bolster their claim by relying on demonstration data that the Census Bureau specifically tuned to *amplify* the infusion of noise so that it could work with its data users to identify and mitigate issues in its various algorithms. But Plaintiffs acknowledge that the Census Bureau will release more-realistic demonstration data later this month. And, as Defendants explain below, those data—which will more-closely resemble the final redistricting data products—will be quite accurate. Plaintiffs nevertheless argue that *any* application of differential privacy will violate the Census Act on the grounds that the resulting data products would not constitute “tabulations of population.” But that argument is belied by the Census Act itself—as well as by Plaintiffs, who themselves refer to the Bureau’s forthcoming redistricting data products in their brief as tabulations of population.

The relief Plaintiffs seek also raises significant concerns. If this Court were to enjoin the use of differential privacy, the Bureau would still need to impose some form of disclosure avoidance. Plaintiffs suggest that the Bureau could use its ineffective 2010 disclosure-avoidance methodology for this year's census. But as explained below, any feasible alternative solution would result in far-less-accurate data and would take months to implement, at a minimum.

Though Plaintiffs ask that the Court prolong the extant delay, they also demand that Defendants produce the redistricting data now. But the redistricting data set does not yet exist, and will likely not come into existence in any form until late August, as the data are still being processed. To the extent that Defendants can produce the redistricting data earlier, they will do so. But any Order from this Court must take into account not only Plaintiffs' desires for the prompt publication of redistricting data, but also the reality that events beyond the Census Bureau's control have delayed the creation and production of those data products.

\* \* \*

The decennial census is an extremely complicated endeavor. It is steered by expert scientists, statisticians, and systems engineers. It is the type of process that should be managed by subject-matter experts ultimately accountable to the elected Executive. "There is no basis for the judiciary to inject itself into this sensitive political controversy and seize for itself the decision to reevaluate the competing concerns between [census] accuracy and speed." *Nat'l Urban League v. Ross*, 977 F.3d 698, 713 (9th Cir. 2020) (Bumattay, J., dissenting from denial of administrative stay), *stay granted*, 141 S. Ct. 18 (2020). The same principle applies here: the Secretary of Commerce and the Census Bureau — not Plaintiffs or this Court — are best positioned to balance accuracy, confidentiality, and speed. Plaintiffs' motion and petition should be denied.

## BACKGROUND

### A. The Decennial Census

“The Constitution requires an ‘actual Enumeration’ of the population every 10 years and vests Congress with the authority to conduct that census ‘in such Manner as they shall by Law direct.’” *Wisconsin v. City of New York*, 517 U.S. 1, 5 (1996) (quoting U.S. Const. art. I, § 2, cl. 3). Congress, in turn, “has delegated to the Secretary of the Department of Commerce the responsibility to take ‘a decennial census of [the] population . . . in such form and content as he may determine.’” *Id.* (quoting 13 U.S.C. § 141(a)). “The Secretary is assisted in the performance of that responsibility by the Bureau of the Census and its head, the Director of the Census.” *Id.* (citing 13 U.S.C. §§ 2, 21).

“The Constitution provides that the results of the census shall be used to apportion the Members of the House of Representatives among the States.” *Id.* And “[b]ecause the Constitution provides that the number of Representatives apportioned to each State determines in part the allocation to each State of votes for the election of the President, the decennial census also affects the allocation of members of the electoral college.” *Id.* “[C]ensus data also have important consequences not delineated in the Constitution: The Federal Government considers census data in dispensing funds through federal programs to the States, and the States use the results in drawing intrastate political districts.” *Id.* at 5–6.

Today, the decennial census is a 15.6-billion-dollar operation, designed to count over 330 million people across 3.8 million square miles. *See* Declaration of Michael Thieme ¶¶ 4–5. And it necessarily requires the cooperation of the American public. For the 2020 census, the Census Bureau spent hundreds of millions of dollars to encourage the country to respond to the census, *see, e.g., id.* ¶ 12, and hundreds of thousands census field staff fanned out across the country to follow up on nonresponding addresses, *see id.* ¶¶ 4, 19–28.

“Although each [decennial census] was designed with the goal of accomplishing an ‘actual Enumeration’ of the population, no census is recognized as having been wholly successful in achieving that goal.” *Wisconsin*, 517 U.S. at 6. As a massive, human-driven operation, the census is, almost by definition, imperfect, despite the monumental efforts of the Census Bureau staff who strive to “count everyone living in the country once, only once, and in the right place.” Thieme Decl. ¶ 3. “Persons who should have been counted are not counted at all or are counted at the wrong location; persons who should not have been counted (whether because they died before or were born after the decennial census date, because they were not a resident of the country, or because they did not exist) are counted; and persons who should have been counted only once are counted twice.” *Wisconsin*, 517 U.S. at 6. As a result, census data “may be as accurate as such immense undertakings can be, but they are inherently less than absolutely accurate.” *Gaffney v. Cummings*, 412 U.S. 735, 745 (1973).

#### **B. The Census Act’s Confidentiality Provisions**

“[A]n accurate census,” of course, “depends in large part on public cooperation.” *Baldrige v. Shapiro*, 455 U.S. 345, 354 (1982). But many people chafe at the notion of providing the government with their personal information. Census Bureau research shows that over half of census respondents were at least “somewhat concerned” – with 28% “very concerned” or “extremely concerned” – about the confidentiality of their census responses. Declaration of John M. Abowd ¶ 11. And “[t]hese concerns are even more pronounced in minority populations and represent a major operational challenge to enumerating traditionally hard-to-count populations.” *Id.*

“To stimulate [the public’s] cooperation Congress has provided assurances that information furnished to the Secretary by individuals is to be treated as confidential.” *Baldrige*, 455 U.S. at 354 (citing 13 U.S.C. §§ 8(b), 9(a)). In particular, sections 8 and 9 of the Census Act provide in part that: (i) “the Secretary [of Commerce] may furnish copies

of tabulations and other statistical materials which do *not* disclose the information reported by, or on behalf of, any particular respondent,” 13 U.S.C. § 8(b) (emphasis added); and (ii) Defendants, and their officers and employees, may not “make *any* publication whereby the data furnished by any particular establishment or individual under this title can be identified,” 13 U.S.C. §§ 9(a), (a)(2) (emphasis added). Indeed, the Census Act provides that Census Bureau staff that publish information protected by § 9 “shall be” subject to fines “or imprisoned not more than 5 years, or both.” 13 U.S.C. § 214. In short, “§ 8(b) and § 9(a) of the Census Act embody explicit congressional intent to preclude *all* disclosure of raw census data reported by or on behalf of individuals.” *Baldrige*, 455 U.S. at 361 (emphasis added).

**C. The Rise of Computing Power and Its Implications for Confidentiality**

In past decennial censuses, the Census Bureau protected the confidentiality of the released data by using disclosure-avoidance mechanisms such as suppression (*i.e.*, withholding data) and, in later censuses, data-swapping (*i.e.*, where certain characteristics of a number of households are swapped with those of other households as paired by a matching algorithm). *Abowd Decl.* ¶¶ 23–25. The 2010 decennial census employed data-swapping as its primary disclosure-avoidance mechanism, and the Census Bureau’s data-swapping methodology kept the total population and total-voting-age population constant for each census block, the smallest level of census geography. *Id.* ¶ 25. This method of disclosure avoidance was considered sufficient at the time. *See id.* ¶¶ 26, 49.

That is no longer the case. It has long been known that purportedly de-identified, aggregated data may be “reconstructed” by a series of mathematical algorithms, though such attacks had been constrained by the limits of available computational power. In one famous example, Professor Latanya Sweeney revealed in 1997 that she had re-identified then-Massachusetts Governor William Weld’s medical records in a purportedly de-identified public database. *See id.* ¶ 27. And as computing power becomes cheaper, more

plentiful, and more accessible as it moves to the cloud, re-identification attacks have increased, and have targeted increasingly large datasets. One recent article recounted re-identification attacks on supposedly de-identified datasets as varied as German internet browsing histories, Australian medical records, New York City taxi trajectories, and London bike-sharing trips. *See* Luc Rocher *et al.*, “Estimating the success of re-identifications in incomplete datasets using generative models,” *Nature Communications* (2019), available [here](#); *see also* Abowd Decl. ¶¶ 33–36 (collecting other examples).

The decennial census is not immune to these trends. Following the 2010 census, the Census Bureau published *over 150 billion independent statistics* about the characteristics of the 308,745,538 persons enumerated in the census. Abowd Decl. ¶ 18. The Census Bureau thus conducted its own reconstruction experiment based on just 6.2 billion of those statistics. The Bureau’s simulated attack precisely reconstructed approximately 46% of the 308,745,538 records with their exact race, ethnicity, sex, and age—and more than 70% of the reconstructed records had exact race, ethnicity, and sex, and were within one year of actual age. *See* Abowd Decl. App’x B ¶¶ 5–7.

The Census Bureau then attempted a re-identification experiment using commercially available databases, and was able to successfully re-identify about 52 million individuals—roughly 17% of the people enumerated in the 2010 census. *See id.* ¶¶ 22–23; Abowd Decl. ¶ 38. And if an attacker had access to data better than the third-party data used in the Census Bureau’s simulation, as many as 179 million people could correctly be re-identified. *See* Abowd Decl. App’x B ¶¶ 24; Abowd Decl. ¶ 38. Although Dr. Abowd had in 2018 described the re-identification risk as “small,” he retracted that tentative conclusion at the February 16, 2019, session of the American Association for the Advancement of Science. *See* Abowd Decl. ¶ 83.

This serious reconstruction and re-identification vulnerability has been confirmed by the JASON group, which Plaintiffs describe as “an independent group of scientists and engineers from whom the Census Bureau has sought third-party review,” and on

whose work Plaintiffs rely. Pls. Mot., Doc. 3 (“Mot.”) at 31. The JASON group explained—in a publication that Plaintiffs repeatedly cited to the Court, *see* Mot. 13 & n.24, 29 & n.57, 31, 32 & nn.58–59—that, in its view, “Census has convincingly demonstrated the existence of a vulnerability that census respondents can be re-identified through the process of reconstructing microdata from the decennial census tabular data and linking that data to databases containing similar information that can identify the respondent.” *See generally* JASON, *Formal Privacy Methods for the 2020 Census* (Apr. 2020) at 89, available [here](#). The JASON group summarized its findings on this point as:

- The Census has demonstrated the re-identification of individuals using the published 2010 census tables.
- Approaches to disclosure avoidance such as swapping and top and bottom coding applied at the level used in the 2010 census are insufficient to prevent re-identification given the ability to perform database reconstruction and the availability of external data.

*Id.* at 6; *accord id.* at 93–94. In short, as Dr. Abowd explains, data produced by the 2010 disclosure-avoidance mechanism would be “vulnerable to reconstruction and re-identification attacks because of the parameters of the swapping mechanism’s 2010 implementation: an overall insufficient level of noise, the invariants preserved without noise, and the geographic and demographic detail of the published summary data.” Abowd Decl. ¶ 39. As such, “[t]he Census Bureau can no longer rely on the swapping implementation used in 2010 if it is to meet its obligations to protect respondent confidentiality.” *Id.*; *see generally id.* ¶¶ 41–43, 50–51.

#### **D. Differential Privacy**

At a fundamental level, all disclosure-avoidance methodologies have a necessary impact on the availability and accuracy of the resulting data. That is how confidentiality is protected. Data-swapping, for example, injects noise into the census redistricting data by swapping certain characteristics between a subset of households. *See* Abowd Decl.

¶¶ 25. But data-swapping—as demonstrated by the Census Bureau and corroborated by the JASON group—is susceptible to database reconstruction attacks. *See id.* ¶¶ 26, 39. And the precise data-swapping methodology used is necessarily opaque, so as to better protect the confidentiality of the data. As Dr. Abowd explains, “[i]mplementation parameters for these legacy disclosure avoidance methods, especially swapping rates, are often some of the most tightly guarded secrets that the Census Bureau protects.” Abowd Decl. ¶ 62.

Given the now-demonstrable flaws with the disclosure-avoidance methodologies used in the 2010 decennial census, “a swapping mechanism that targets vulnerable households for swapping would require significantly higher rates of swapping than were used in 2010 to protect against a reconstruction attack.” *Id.* ¶ 42. And utilizing such higher swapping rates would “have a significant, detrimental impact on data quality.” *Id.* Moreover, “[i]mplementing swapping in 2020 would also require abandoning the total population and voting-age population invariants that were used in 2010” for two reasons: (i) it would be “impossible to find enough paired households with the same number of persons and adults without searching well outside the neighborhood of the original household”; and (ii) “holding the total and adult populations invariant gives the attacker a huge reconstruction advantage—exact record counts in each block for persons and adults”—and that advantage “vastly improves the accuracy of the reconstructed data.” *Id.* But “[i]nternal experiments . . . confirmed that increasing the swap rate from the level used in 2010 and removing the invariants on block-level population counts (to permit the increased level of swapping and protect against reconstruction attacks) would render the resulting data unusable for most data users.” *Id.*

Nor is data suppression a viable option. “While the Census Bureau could use suppression to protect from a reconstruction attack, the resulting data would be only available at a very high level of generality.” *Id.* ¶ 43. “Today’s data users, including redistricters, rely on detailed block and tract-level data, which would not be available for



many areas if the Census were to return to suppression to protect against modern attacks.” *Id.*

Ultimately, the Census Bureau’s Data Stewardship Executive Policy Committee (DSEP) determined that neither swapping nor suppression would allow the Census Bureau “to produce high quality statistics from the decennial census while also protecting the confidentiality of respondents’ census records” as required by the Census Act. *Id.* ¶ 46; *see also id.* ¶ 51 (“[T]o achieve the necessary level of privacy protection, both enhanced data swapping and suppression had severely deleterious effects on data quality and availability.”).

This led the Census Bureau to differential privacy, “[t]he best disclosure avoidance option that offers a solution capable of addressing the new risks of reconstruction-abetted re-identification attacks, while preserving the fitness-for-use of the resulting data for the important governmental and societal uses of census data.” *Id.* ¶ 47. Differential privacy is used by major private-sector technology firms, and the Census Bureau has been using differential privacy to protect certain of its statistical products since 2008. *See id.* ¶ 45.

“Differential privacy, first developed in 2006, is a framework for quantifying the precise disclosure risk associated with each incremental release from a confidential data source.” *Id.* ¶ 44. This framework allows “the Census Bureau to quantify the precise amount of statistical noise required to protect privacy.” *Id.* “This precision allows the Census [Bureau] to calibrate and allocate precise amounts of statistical noise in a way that protects privacy while maintaining the overall statistical validity of the data.” *Id.* The amount of noise injected is determined by a measure known as the privacy-loss budget (PLB) or the “epsilon.” Michael Hawes, U.S. Census Bureau, “Differential Privacy and the 2020 Decennial Census” (Mar. 5, 2020), at 18, available [here](#). Setting epsilon to zero would result in perfect privacy but useless data, and setting the epsilon to infinity would result in perfect accuracy, but would result in releasing data in fully identifiable form. *Id.*

The advantages of differential privacy are myriad. *See, e.g.*, Simson L. Garfinkel, U.S. Census Bureau, *Modernizing Disclosure Avoidance* (Sept. 15, 2017) at 10, available [here](#). Those advantages include protection against database reconstruction attacks and privacy guarantees that do not depend on the availability of external data. *See id.* It can do so while still producing highly accurate data. Abowd Decl. ¶ 54. And, as will be implemented by the Census Bureau, the accuracy of the data increases, not decreases, as census geographies increase in size. *See id.* ¶ 56.

Moreover, differential privacy can be tuned to determine the optimal setting whereby the privacy of confidential data can be reasonably assured, yet the resulting data will be fit for redistricting and other uses. *See id.* ¶¶ 52, 54, 59. The Bureau’s “empirical analysis showed that differential privacy offered the most efficient trade-off between privacy and accuracy—[its] calculations showed that the efficiency of differential privacy dominated traditional methods.” *Id.* ¶ 41. “In other words, regardless of the level of desired confidentiality, differential privacy will always produce more accurate data than the alternative traditional methods considered by the Census Bureau.” *Id.*

Differential privacy also allows for unprecedented transparency. “The Census Bureau has submitted its differential privacy mechanisms, programming code, and system architecture to thorough outside peer review.” Abowd Decl. ¶ 62. The Bureau has “also committed to publicly releasing the entire production code base and full suite of implementation settings and parameters.” *Id.* Whereas swapping techniques “must be implemented in a ‘black box,’” to protect the resulting data, differential privacy, by contrast, “does not rely on the obfuscation of its implementation as a means of protecting the data.” *Id.* “The Census Bureau’s transparency will allow any interested party to review exactly how the algorithm was applied to the 2020 Census data, and to independently verify that there was no improper or partisan manipulation of the data.” *Id.*

And the Census Bureau has aimed to tune the disclosure-avoidance algorithms, and will tune the privacy-loss budget, in the public eye. *See generally id.* ¶¶ 57–62. In

October 2019 and throughout 2020, the Census Bureau publicly released “demonstration data.” See U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), available [here](#). Exactly as designed, these public releases resulted in “extensive actionable feedback from the data user community,” which “has informed ongoing [disclosure-avoidance] system improvements and design changes.” *Id.* During this iterative process, the Census Bureau “used a lower privacy-loss budget than [it] anticipate[s] using for the final 2020 Census data – that is, these demonstration data were purposefully ‘tuned’ to privacy and not ‘tuned’ for producing highly accurate redistricting data.” Abowd Decl. ¶ 61. The Bureau did so in order “to home in on the elements of the algorithm that were causing systemic distortions that needed to be addressed.” U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), available [here](#). This decision “meant that the resulting [demonstration] data would have substantially more noise (error) than should be expected in the final 2020 Census data products,” but it “unfortunately led some of our data users to expect comparable amounts of noise in the final 2020 Census data.” *Id.*

Fortunately, that will not be the case. By keeping the privacy-loss budget roughly constant in the demonstration data to date, the Census Bureau has been able to improve the post-processing algorithms and mitigate post-processing errors. See U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 3, 2021), available [here](#).<sup>1</sup> For example, “the Census Bureau has identified and corrected the algorithmic sources of [certain] distortions,” and “any residual impact of the types of systematic bias observed in the early

---

<sup>1</sup> The *amicus* States prove this point. They note, for example, that Utah “analyzed the 2010 demonstration data, comparing it with the previously received 2010 redistricting data and sent its findings to the Census Bureau.” Doc. 40 at 2. And they acknowledge that this iterative process worked: Utah acknowledges that it “saw an improvement from the October 2019 to the November 2020 demonstration data,” *id.* at 3, though they incorrectly attribute that improvement to modifications in the privacy-loss budget. See *id.*

demonstration data will be negligible and well within the normal variance and total error typical for a census.” Abowd Decl. ¶ 67.

And with those algorithmic improvements in place, the Census Bureau moved to tuning the privacy-loss budget. “On March 25, 2021, DSEP approved the privacy-loss budget to be used for the next demonstration product. This privacy-loss budget reflects empirical analysis of over 600 full-scale runs of the Disclosure Avoidance System using 2010 Census data.” Abowd Decl. ¶ 70. “The Census [Bureau] evaluated these experimental runs using accuracy and fitness-for-use criteria for the redistricting use case informed by the extensive feedback we have received from the redistricting community and the Civil Rights Division at the U.S. Department of Justice.” *Id.*

The Census Bureau intends to release the next set of demonstration data by April 30, 2021. *See U.S. Census Bureau, 2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), available [here](#). This set of data employs a higher privacy-loss budget, tuned for accuracy, “that better approximates the final privacy-loss budget that will likely be selected for the redistricting data product.” Abowd Decl. ¶ 69. “These new demonstration data will also reflect system design changes that have been made since the last demonstration data release, along with tuning and optimization of the system that have been done specifically to prioritize population count accuracy and the ability to identify majority-minority districts.” *Id.*

“The next iteration of demonstration data will establish that differential privacy protections can produce extremely accurate redistricting data.” Abowd Decl. ¶ 54. In the upcoming release of demonstration data:

- “Total populations for counties have an average error of +/- 5 persons . . . as noise from differential privacy” (an error rate of about 0.04% of the counties’ population). Compare that level of precision with the “average county-level” estimated uncertainty inherent in census counts, which “is +/- 960 persons (averaging 1.6% of the county census counts).” *Id.*

- “At the block level the differentially private data have an average population error of +/- 3 persons” which is also more precise than “the simulated error inherent in the census which puts the average error uncertainty of block population counts at +/- 6 people.” *Id.*
- “In the April 2021 Demonstration Data Product, Congressional districts as drawn in 2010 [nationwide] have a mean absolute percentage error of 0.06%.” *Id.* ¶ 56.
- “Even for state legislative districts, which had average sizes of 159,000 (upper chambers) and 64,000 (lower chamber[s]), the mean absolute percentage errors are 0.09% (upper chambers) and 0.16% (lower chambers), respectively. Such errors are trivial and imply that the difference between districts drawn from the April 2021 Demonstration Data Product and those drawn from the original 2010 P.L. 94-171 Redistricting Data Summary File would be statistically and practically imperceptible.” *Id.*
- “The April 2021 demonstration data show no meaningful bias in the statistics for racial and ethnic minorities even in very small population geographies like Federal American Indian Reservations.” *Id.* ¶ 55 (emphasis omitted). “The data permit assessment of the largest OMB-designated race and ethnicity group in each geography – the classification used by the Department of Justice for Voting Rights Act scrutiny – with a precision of 99.5% confidence in variations of +/- 5 percentage points for off-spine geographies as small as 500 persons, approximately the minimum voting district size in the redistricting plans that the Department of Justice provided as examples.” *Id.*

In sum, the demonstration data that will be released later this month will demonstrate that the differential-privacy algorithm, “when properly tuned, ensures that redistricters

can remain confident in the accuracy of the population counts and demographic characteristics of the voting districts they draw, despite the noise in the individual building blocks.” *Id.* ¶ 56 (emphasis omitted).

Data-users will have at least four weeks to review the next set of demonstration data, perform their analyses, and submit feedback. *See* U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), available [here](#). In early June, DSEP will set the final privacy-loss budget and production parameters for the redistricting data product. *See id.* Applying differential privacy to the redistricting data will take roughly three weeks—“similar to the period required to implement disclosure avoidance in prior censuses”—and “is not the cause of the delay in the delivery of the redistricting data.” *Abowd Decl.* ¶ 72. In fact, “the disclosure avoidance procedures completed in the 2010 census processing took 27 days--or nearly *four* weeks.” *Thieme Decl.* ¶ 71 (emphasis added).

To the contrary, shifting disclosure-avoidance methodologies now is all but guaranteed to cause further delay—and “[t]he effect on the schedule for delivering redistricting data would be substantial.” *Abowd Decl.* ¶¶ 84–85. “[U]nder all scenarios the delay would be *multiple months*.” *Id.* ¶ 85 (emphasis added). “This delay is unavoidable because the Census Bureau would need to develop and test new systems and software, then use them in production and subject the results to expert subject matter review prior to production of data.” *Id.*

Because the 2010 census data are vulnerable to a database reconstruction attack, “the Census Bureau cannot simply repeat the swapping protocols from the 2010 census, but rather would be forced to fashion appropriate levels of protection”—and “[u]sing an appropriate level of protection for either suppression or swapping would produce far less accurate data than would differential privacy.” *Id.* ¶ 87. And even if the Census Bureau were “ordered to repeat exactly what was done in 2010 (despite the serious risks to privacy the Census has identified),” the Bureau “could not simply ‘flip a switch’ and

revert to the prior methodology.” *Id.* ¶ 86. “The 2020 Census’s system architecture is completely different than that used in the 2010 Census, and it is thus not possible to simply ‘plug in’ the disclosure-avoidance system used in 2010.” *Id.* “Instead,” the Bureau “would need to conduct the requisite software development and testing.” *Id.*

Simply put, it is not practical at this late hour to change the disclosure-avoidance system’s methodologies. Such decisions “are highly technical and can have unanticipated consequences.” *Id.* ¶ 88. “While [the Census Bureau] cannot predict the full impact of any change, there is a danger than any change would have cascading effects on data accuracy and privacy, making race and ethnicity data, along with age data, substantially less accurate.” *Id.* And “[a]ny sort of change in the basic methodology would be minimally tested and would not have the benefit of any input from the user community.” *Id.*

#### **E. The Census Bureau’s Delivery of Redistricting Data**

As explained above, the 2020 Census has been a massive undertaking. While the Bureau has done everything in its power to complete the census as expeditiously as possible, the COVID-19 pandemic has resulted in some unavoidable delay. The original plan was for the Census Bureau to begin in-person operations (called Nonresponse Followup or NRFU) in May 2020, but it was forced to suspend those operations for months due to the pandemic. Thieme Decl. ¶ 30. By the time the Census Bureau entered the field in earnest three months later, it did so during a perfect storm of natural disasters and civil unrest. *Id.* ¶ 33. “Devastating hurricanes in the Gulf Coast area . . . limited and slowed the Census Bureau’s ability to conduct NRFU operations.” *Id.* In “large areas of the West Coast, field operations were hampered by conflagrations that caused health alerts due to fire and smoke.” *Id.* And “in cities across the country,” civil unrest made the already-difficult enumeration even harder. *Id.*

Making matters worse, the Secretary and the Census Bureau were under a statutory directive to report the census results to the President by December 31, 2020 so that he could timely submit them to Congress for reapportionment of the House. *See* 13 U.S.C.

§ 141(b); 2 U.S.C. § 2a. And although the Secretary had asked for an extension of these statutory deadlines, Congress did not oblige. Thieme Decl. ¶ 35. So the Census Bureau again adjusted its operations in an attempt to meet the statutory deadlines. *Id.* ¶ 36. But that adjustment led to the intervention of another Branch: the Judiciary. After a court-ordered preliminary injunction forced the Census Bureau to remain in the field, an emergency Supreme Court ruling stayed that injunction and allowed the Census Bureau to conclude field operations in mid-October 2020, having resolved 99.9% of all housing units in the process. *See Ross v. Nat'l Urban League*, 141 S. Ct. 18 (2020); Thieme Decl. ¶ 36.

But collecting responses through completed questionnaires and in-person field work is not the end of the story – the Census Bureau must then summarize the individual and household data that it collected into usable, high-quality tabulations. Thieme Decl. ¶¶ 37–83. Although creating such tabulations may appear easy, it is not. The Census Bureau must integrate data from different enumeration methods used across the country, identify any issues or inconsistencies that arise, rectify them, and produce tabulations that will guide the country for the next ten years, all without compromising its statutory mandate to maintain the confidentiality of census responses. 13 U.S.C. §§ 8, 9; Thieme Decl. ¶¶ 53–59 (describing how administrative records are incorporated and data are reconciled to produce the Census Unedited File); *id.* ¶¶ 60–64 (describing how the federally affiliated overseas population is incorporated into the data to produce apportionment numbers); *id.* ¶¶ 65–70 (describing the iterative process for compiling detailed information such as race, ethnicity, and age to produce the Census Edited File); *id.* ¶¶ 71–74 (describing the process for applying the Census Bureau’s disclosure-avoidance methodology); *id.* ¶¶ 75–78 (describing the process for generating usable data files).

Even working with all possible dispatch, the Census Bureau was not able to meet its December 31, 2020 statutory deadline for reporting apportionment numbers. Due to the difficulties encountered during data collection and issues that arose during the processing phase, the Census Bureau projects that it will not complete apportionment counts



until April 30, 2021. Thieme Decl. ¶ 37. Another court and other parties have even relied upon Defendants' representation that "the Census Bureau will not under any circumstances report the results of the 2020 Census . . . before April 16, 2021." *Nat'l Urban League v. Raimondo*, No. 20-cv-05799, ECF Nos. 465 & 467 (N.D. Cal. Feb. 3, 2021).

The delay in producing apportionment data also means the Secretary and the Census Bureau have missed the statutory deadline (March 31, 2021) to submit census-based redistricting data to the States. 13 U.S.C. § 141(c). This was not a secret. In a February 12, 2021 Press Release, the Census Bureau explained that "it will deliver the [ ] redistricting data to all states by Sept. 30, 2021" because "COVID-19-related delays and prioritizing the delivery of the apportionment results delayed the Census Bureau's original plan to deliver the redistricting data to the states by March 31, 2021." *Census Bureau Statement on Redistricting Data Timeline*, U.S. Census Bureau (Feb. 12, 2021), available [here](#).

That announcement was not for the Census Bureau's benefit, but for States that use census-based redistricting data to draw their congressional or state election districts. While no federal law requires the use of census data for this purpose, the data are generally utilized as the gold standard, including by the Department of Justice, which uses such data for enforcement of the Voting Rights Act. Declaration of James Whitehorne ¶ 4. That's why States generally use census data for redistricting. And many of those States make up the 27 States that are bound by their own laws to redistrict in 2021. *See 2020 Census Delays and the Impact on Redistricting*, National Conference of State Legislatures (last visited Apr. 11, 2021), available [here](#). That has led some States under self-imposed redistricting pressure to find workable solutions. In New Jersey, for example, voters approved a constitutional amendment that allowed the State to use previous district maps until the new maps are in effect for the 2023 elections. *See Whitehorne Decl. ¶ 7*; N.J. Const. art. IV, § 3, ¶ 4. And in California, the state legislature sought and obtained at least a four-month delay of the redistricting deadlines from the California Supreme Court. *Legislature of the State of Cal. v. Padilla*, 469 P.3d 405, 413 (Cal. 2020);

Whitehorne Decl. ¶ 7. These States—and many others—gathered information from the Census Bureau and found a way to remedy their own redistricting issues. Whitehorne Decl. ¶¶ 7-8.

Alabama is not one of those States. Instead, Alabama now seeks redistricting data that does not exist by a statutory deadline that is impossible to meet. Whitehorne Decl. ¶¶ 14-16. Defendants oppose that request.

## ARGUMENT

### I. PLAINTIFFS LACK STANDING.

“The doctrine of standing is an essential and unchanging part of the case-or-controversy requirement embodied in Article III of the Constitution.” *Flat Creek Transp., LLC v. Fed. Motor Carrier Safety Admin.*, 923 F.3d 1295, 1300 (11th Cir. 2019).<sup>2</sup> “In the absence of standing, a court is not free to opine in an advisory capacity about the merits of a plaintiff’s claims, and the court is powerless to continue.” *Aaron Private Clinic Mgmt. LLC v. Berry*, 912 F.3d 1330, 1335 (11th Cir. 2019).

“The irreducible constitutional minimum of standing requires a plaintiff to show that he (1) suffered an injury in fact, (2) that is fairly traceable to the challenged conduct of the defendant, and (3) that is likely to be redressed by a favorable judicial decision. *Flat Creek Transp., LLC*, 923 F.3d at 1300. “[A]s the part[ies] invoking federal jurisdiction,” Plaintiffs “bear[] the burden of establishing these elements.” *Id.* “And because standing doctrine is intended to confine the federal courts to a properly judicial role,” those courts must “take seriously the requirement that a plaintiff *clearly* demonstrate each requirement.” *Id.* (emphasis added). “If the plaintiff fails to meet its burden, this court lacks the power to create jurisdiction by embellishing a deficient allegation of injury.” *Aaron Private Clinic Mgmt. LLC*, 912 F.3d at 1336.

---

<sup>2</sup> Unless expressly included, all citations and internal quotation and alteration marks have been omitted.

Plaintiffs have not demonstrated—let alone “clearly” demonstrated—any of the three necessary standing elements. Accordingly, their motion should be denied.

**A. Plaintiffs Have Not Sustained Any Injuries-in-Fact**

An “injury in fact” is “the invasion of a judicially cognizable interest that is concrete and particularized and actual and imminent.” *Corbett v. Transp. Sec. Admin.*, 930 F.3d 1225, 1228 (11th Cir. 2019). Plaintiffs have not demonstrated that any of them have been injured or will imminently be injured, either by the application of differential privacy, or by the delay in producing the redistricting data.

**1. Plaintiffs Are Not Injured by Differential Privacy**

Plaintiffs assert five forms of injury-in-fact in connection with their differential-privacy claims. None has merit.

**a. Informational Injury**

Asserting a supposed informational injury, Plaintiffs argue that Alabama is statutorily entitled to “tabulations of population” under 13 U.S.C. § 141(c), *see* Mot. 29–33; Compl. ¶¶ 133–140—and that is precisely what the Secretary will provide to the State. Plaintiffs acknowledge that the term “‘tabulate’ has long been understood to mean ‘[t]o put or arrange in a tabular, systemic, or condensed form.’” Mot. 29 n.57 (quoting *The Random House College Dictionary* 1337 (revised ed. 1975)). It follows that a “tabulation” is the arrangement of data in such form. And Plaintiffs do not dispute that the Secretary will provide to the State data in such an arranged form. Hence, Alabama will receive “tabulations.”

One need only review Plaintiffs’ brief to confirm this fact. Plaintiffs contend that the “*tabulations*” “will be intentionally scrambled.” *Id.* at 2. They allege that they will suffer harm from supposedly “flawed *tabulations*.” *Id.* at 4 (emphasis added). They express concern about supposedly “false *tabulations*.” *Id.* at 27 (emphasis added). They argue that “Defendants plan to provide the State with inaccurate *tabulations*.” *Id.* at 34 (emphasis added). And they contemplate what might happen if “both *tabulations*” —*i.e.*,

tabulations with and without the application of differential privacy – “can be released.” *Id.* at 55 (emphasis added). Plaintiffs may not agree with the methodology that will underlie the Secretary’s tabulations, but Plaintiffs readily acknowledge that they are, in fact, tabulations.

These tabulations will further constitute the “tabulations of population” contemplated in § 141(c). Plaintiffs do not contend that the Secretary will simply invent population numbers. Rather, to ensure compliance with the confidentiality requirements imposed by Congress, *see* 13 U.S.C. §§ 8 & 9, the Census Bureau will inject slight statistical “noise” into the sub-state population counts. *See, e.g.,* Abowd Decl. ¶¶ 54, 69. But that process hardly renders the resulting data something other than “tabulations of population.”

Again, Plaintiffs themselves prove the point. They claim that the Secretary will, in their view, “provide the States purposefully flawed *population tabulations*.” Mot. 1–2 (emphasis added). They contend that “[i]f the Census Bureau uses differential privacy, the *population tabulations* it reports to States for redistricting will be inaccurate.” *Id.* at 24 (emphasis added); *accord id.* at 25. They represent that “[t]he Court will be unable to remedy” supposed “harms if Defendants deliver *population tabulations* infected by differential privacy.” *Id.* at 27 (emphasis added). They argue about what might happen “once the skewed *population tabulations* are delivered.” *Id.* at 51 (emphasis added). And they talk about losing funding “if the *population tabulations* are inaccurate.” *Id.* at 52 (emphasis added); *see also, e.g., id.* at 4 (characterizing differential privacy as “a ‘statistical method’ used ‘to determine the population for purposes of . . . redistricting’”); Pls. Reply, Doc. 25, at 4 (“Challenges to statistical methods that ‘determine the population for purposes of the apportionment or redistricting’ must be heard by a three-judge court.”) (emphasis omitted). But they admit that the tabulations that the Secretary will deliver are, in fact, “tabulations of population.”

In an effort to call into question future population tabulations, Plaintiffs point to their experts' analysis of the Census Bureau's releases of demonstration data. *See generally* Mot. 18–24. Yet Plaintiffs acknowledge that “[f]or the demonstration data products, the Census Bureau set a more conservative privacy-loss budget than it expects will be set for the 2020 census—meaning that the demonstration data will have more ‘noise (error) than should be expected in the final 2020 Census data products.’” *Id.* at 18 (quoting U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021)).

In fact, the Census Bureau explained that it maintained this conservative privacy-loss budget—even though doing so “meant that the resulting data would have *substantially* more noise (error) than should be expected in the final 2020 Census data products”—so the Bureau and its data users could “home in on the elements of the algorithm that were causing systemic distortions that needed to be addressed.” U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), available [here](#) (emphasis added). The Census Bureau is planning to release the next set of demonstration data on April 30, 2021. *Id.*; *see* Mot. 49 (acknowledging same). That demonstration data: (i) “will feature a higher [privacy-loss budget] and system parameter optimization informed by the hundreds of full-scale [disclosure-avoidance system] experimental runs [the Bureau has] been performing over the last several months”; (ii) “will more closely approximate the expected accuracy and fitness-for-use of the final 2020 Census redistricting data product”; and (iii) “will enable [the Bureau’s] data users to provide critical fitness-for-use analyses” and to “submit feedback and recommendations prior to” the Bureau’s Data Stewardship Executive Policy Committee’s decision that will set the final privacy-loss budget in June. U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), available [here](#). Indeed, the average population error in the forthcoming April 30 demonstration data falls well within the estimated uncertainty inherent in the census. *See* Abowd Decl. ¶ 54; *see supra* Background Part D.

Because Plaintiffs do not know how the privacy-loss budget will ultimately be set, or how that future budget will affect the redistricting data, their challenge to differential privacy is facial in nature. Plaintiffs concede as much. Admitting that the final redistricting data will be subject to less noise than the demonstration data to date, Plaintiffs argue that “no matter where the epsilon value is set,” the redistricting data “will just be less wrong than the demonstration numbers were,” and that “*any* application of differential privacy will produce erroneous numbers.” Mot. 18, 35 (emphasis added). In other words, Plaintiffs acknowledge their burden on this facial challenge: they “must establish that no set of circumstances exists under which” the application of differential privacy “would be valid.” *Reno v. Flores*, 507 U.S. 292, 301 (1993) (no-set-of-circumstances test applies to “both the constitutional challenges . . . and the statutory challenge”); *accord, e.g., Associated Builders & Contractors of Tex., Inc. v. Nat’l Labor Relations Bd.*, 826 F.3d 215, 220 (5th Cir. 2016); *Scherer v. U.S. Forest Serv.*, 653 F.3d 1241, 1243 (10th Cir. 2011) (Gorsuch, J.); *Sherley v. Sebelius*, 644 F.3d 388, 397 (D.C. Cir. 2011). And “[t]his heavy burden makes such an attack the most difficult challenge to mount successfully.” *Doe v. Kearney*, 329 F.3d 1286, 1294 (11th Cir. 2003).

Plaintiffs’ effort to satisfy their heavy burden rests on the theory that the “tabulation of total population by States” referenced in § 141(b) is equivalent to the “actual population counts for States,” and “[i]t follows that the ‘tabulations of population’ referenced in subsection 141(c) must also be the actual population counts.” Mot. 30. But nothing in § 141(b) suggests that the term “tabulation” contemplates any particular methodology. The methodology used to determine the apportionment counts stems from the *Constitution*, which requires that the apportionment of Representatives be based on an “actual Enumeration.” U.S. Const. art. I, § 2, cl. 3; *see Dep’t of Commerce v. U.S. House of Representatives*, 525 U.S. 316, 346–47 (1999) (Scalia, J., concurring) (“Dictionaries roughly contemporaneous with the ratification of the Constitution demonstrate that an ‘enumeration’

requires an actual counting, and not just an estimation of number.”). Section 141(b) references only “[t]he *tabulation* of total population by States,” 13 U.S.C. § 141(b) (emphasis added), and not, for example, “[t]he *enumeration* of total population by States.” It does not make sense, then, for Plaintiffs to attempt to synonymize “tabulation” with “enumeration.” *Cf. Firststar Bank, N.A. v. Faul*, 253 F.3d 982, 991 (7th Cir. 2001) (noting “the canon that different words within the same statute should, if possible, be given different meanings”). Instead, Congress used the term “tabulation of total population” in § 141(b) to mean exactly what it says—and how Plaintiffs use it repeatedly in their brief, *see supra*: an arrangement of population data for transmission to the President. *Conn. Nat’l Bank v. Germain*, 503 U.S. 249, 253–54 (1992) (“We have stated time and again that courts must presume that a legislature says in a statute what it means and means in a statute what it says there.”). Put simply, Plaintiffs’ invocation of the obvious—that the word “tabulation” appears in both § 141(b) and § 141(c)—is a non sequitur; it proves only that Congress wanted the Secretary to arrange population data for two different distributions.

And even if the term “tabulation” in § 141(b) could be construed to incorporate a particular methodology, the Census Act itself disproves the notion, *contra* Mot. 30, that any such methodology carries over to § 141(c). For example, the data that underlie the § 141(c) tabulations may be based on statistical sampling, whereas the data that underlie the § 141(b) tabulation may not. Section 195 of the Census Act provides that “the Secretary shall, if [s]he considers it feasible, authorize the use of the statistical method known as ‘sampling’ in carrying out the provisions of this title” — “[e]xcept for the determination of population for purposes of apportionment of Representatives.” 13 U.S.C. § 195. So the data that underlie “[t]he tabulation of total population by States . . . as required for the apportionment of Representatives,” § 141(b), cannot be premised on statistical sampling. But § 195 expressly provides that determinations of population for non-apportionment purposes—such as the redistricting data contemplated by § 141(c)—may properly be based on statistical sampling. *See, e.g., Glavin v. Clinton*, 19 F. Supp. 2d 543, 552–53 (E.D.

Va. 1998) (three-judge court) (“[T]he only plausible interpretation of the plain language and structure of the Act is that Section 195 prohibits sampling for apportionment and Section 141 allows it for all other purposes.”), *aff’d sub nom., Dep’t of Commerce v. U.S. House of Representatives*, 525 U.S. 316 (1999). In other words, nothing in the Census Act would preclude the Secretary from both: (i) producing the “tabulation of total population by States . . . as required for the apportionment of Representatives” under § 141(b) based on the actual enumeration; and (ii) developing the sub-state “[t]abulations of population” contemplated by § 141(c) through, say, a hybrid enumeration-and-statistical-sampling protocol.

This point is further borne out by the drafting history of the Census Act. Congress added § 141(c) in December 1975 but did not at that time amend § 195 to carve out the § 141(c) tabulations from § 195’s statistical-sampling authorization. *See* Pub. L. No. 94-171, 89 Stat. 1023 (Dec. 23, 1975). And less than a year later, Congress amended both § 141(c) and § 195. *See* Pub. L. No. 94-521 §§ 7(a) & 10, 90 Stat. 2459 (Oct. 17, 1976). But Congress again declined to carve out the § 141(c) tabulations from § 195’s statistical-sampling authorization. Congress’s intent, as expressed through its legislative decisions and statutory text, is clear: statistical sampling is off limits only when “determin[ing] [the] population for purposes of apportionment of Representatives.” 13 U.S.C. § 195. In every other context—including the redistricting context—statistical sampling is fair game. So the Census Act’s structure and drafting history disproves the thesis central to Plaintiffs’ legal theory: that the data underlying the tabulations contemplated in § 141(c) must be premised on the same methodology as those that underlie the tabulation contemplated in § 141(b). Rather, the Census Act itself demonstrates that the data underlying § 141(b) and § 141(c) may differ in methodology.

Plaintiffs also seem to argue in passing that the *Constitution* somehow obligates Defendants to produce redistricting data through their preferred methodology. Mot. 31. The single case Plaintiffs cite says nothing of the sort, and they quickly back away from



this undeveloped this argument. *See id.* (“At the very least, the constitutional question is raised . . .”). But “[i]t is not enough merely to mention a possible argument in the most skeletal way, leaving the court to do counsel’s work, create the ossature for the argument, and put flesh on its bones.” *United States v. Zannino*, 895 F.2d 1, 17 (1st Cir. 1990); *see Forsberg v. Pefanis*, 634 F. App’x 676, 680 (11th Cir. 2015) (“Pefanis makes two other arguments, both of which he has forfeited by failing to develop them.”). In all events, Plaintiffs are mistaken. “[T]he constitutional purpose of the census” is “to determine the *apportionment* of the Representatives among the States,” *Wisconsin*, 517 U.S. at 20 (emphasis added)—that is, to determine the number of Representatives to which each State is entitled after the decennial census. Though “the States use the [census] results in drawing intrastate political districts,” that “consequence[.]” is “not delineated in the Constitution.” *Id.* at 5–6 (emphasis added); *see also* Departments of Commerce, Justice, and State, The Judiciary, and Related Agencies Appropriations Act, 1998, § 209(a)(2), Pub. L. No. 105–119, 111 Stat. 2440 (1997) (codified at 13 U.S.C. § 141 note) (“1998 Appropriations Act”) (“[T]he *sole* constitutional purpose of the decennial enumeration of the population is the *apportionment* of Representatives in Congress among the several States.”) (emphases added).

Plaintiffs fare no better in attempting to import a judicially enforceable “accuracy” requirement into § 141(c).<sup>3</sup> The decennial enumeration is an attempt to determine the true population of the United States, and “[t]hese figures may be as accurate as such immense undertakings can be.” *Gaffney*, 412 U.S. at 745. But as a matter of reality, census data “are inherently less than absolutely accurate.” *Id.* “Those who know about such

---

<sup>3</sup> *Amica* Professor Bambauer argues that an accuracy requirement can be found in 13 U.S.C. § 181. Doc. 33 at 20–21. Even assuming that Professor Bambauer’s interpretation of § 181 were correct, § 181 expressly concerns certain data produced “[d]uring the *intervals between* each census of population required under section 141.” 13 U.S.C. § 181(a) (emphasis added). It does not relate to the data produced pursuant to § 141(c).

things,” the Supreme Court explained, “recognize this fact.” *Id.* And even if the enumeration could somehow result in a perfect population count, “the well-known restlessness of the American people means that population counts for particular localities are outdated long before they are completed.” *Karcher v. Daggett*, 462 U.S. 725, 732 (1983); *see also, e.g., Gaffney*, 412 U.S. at 745–46 (“[I]t makes little sense to conclude from relatively minor ‘census population’ variations among legislative districts that any person’s vote is being substantially diluted. The ‘population’ of a legislative district is just not that knowable to be used for such refined judgments.”).

In other words, the population counts determined in the decennial census are an approximation within a statistical range of the inherently unknowable population on Census Day. *See* Abowd Decl. ¶ 54. And the Census Bureau expects that the statistical “noise” that the differential-privacy algorithm will inject into those numbers will be measurably within that statistical range. *See id.* ¶¶ 54, 69. And in many cases, the post-differential-privacy population counts will have the effect of being *more* accurate.

For example, say the actual (but inherently unknowable) population of a given census block on Census Day is 50 individuals. The population count as determined by the actual enumeration might nonetheless record only 47 individuals as residing in the census block. But after the differential-privacy algorithm has been applied, the resulting population count increases by one person, *i.e.*, to 48 individuals. Plaintiffs’ legal position is that the post-differential-privacy population count of 48 individuals is illegally inaccurate while the 47-person figure is not – even though the 48-person figure is, in truth, more accurate. Such a result would not make sense.

Moreover, Plaintiffs’ position – that the Census Act incorporates *sub silentio* a judicially enforceable accuracy requirement hiding somewhere in the Census Act’s penumbrae, *see* Mot. 32–33 – is the precise argument adopted by the district court in *National Urban League v. Ross* in enjoining the Secretary’s attempt to comply with the statutory apportionment deadline on the grounds that it was trumped by a supposed “statutory

duty of accuracy.” 489 F. Supp. 3d 939, 982, 994 (N.D. Cal. 2020), *stay denied in part*, 977 F.3d 770 (9th Cir. 2020), *stay granted*, 141 S. Ct. 18 (2020). We know how that ended: with a “rare and exceptional” Supreme Court stay. *Fargo Women’s Health Org. v. Schafer*, 507 U.S. 1013, 1014 (1993) (O’Connor, J., concurring in denial of stay application); *see Ross v. Nat’l Urban League*, 141 S. Ct. 18 (2020). And the Supreme Court granted the government’s requested stay despite the solo dissenter’s position that “respondents [would] suffer substantial injury if the Bureau is permitted to sacrifice accuracy for expediency.” *Nat’l Urban League*, 141 S. Ct. at 21 (Sotomayor, J., dissenting).

“Through the Census Act, Congress has delegated its broad authority over the census to the Secretary.” *Wisconsin*, 517 U.S. at 19. And the Secretary and the Census Bureau—not Plaintiffs or the Court—are best positioned to optimally balance accuracy and confidentiality. Indeed, “there’s one branch Congress has not delegated any census decisions to: the judiciary.” *Nat’l Urban League*, 977 F.3d at 704 (Bumatay, J., dissenting). And just as “[t]here is no basis for the judiciary to inject itself into this sensitive political controversy and seize for itself the decision to reevaluate the competing concerns between accuracy and speed,” *see id.* at 713 (Bumatay, J., dissenting), there is similarly no basis for this Court to inject itself into the Census Bureau’s disclosure-avoidance methodology and seize for itself the decision to reevaluate the competing concerns between accuracy and confidentiality.

In sum, the Secretary will provide to the States redistricting data subject to differential privacy. Those data will be provided in a “tabulation,” and they represent the sub-state population. They are hence “tabulations of population.” 13 U.S.C. § 141(c). Because the Secretary will provide Alabama with “tabulations of population” as afforded to the State in § 141(c), “Defendants’ decision to apply differential privacy will” not “deprive Alabama of information which it is entitled to receive.” *Contra Mot.* 32. Alabama thus suffers no informational injury.

**b. Sovereign Injury**

Plaintiffs argue that the application of differential privacy will injure Alabama by “imped[ing] the State’s sovereign interest in drawing fair districts.” Mot. 33. In fact, Alabama will suffer no such injury for two independent reasons.

*First*, the redistricting data that the Secretary will ultimately produce to Alabama will be perfectly fit for redistricting. As explained above, the redistricting data need not exactly reflect the population counts from the enumeration, and the Census Bureau expects that the noise injected by differential privacy will be less than the estimated uncertainty inherent in the census. *See* Abowd Decl. ¶¶ 54, 69. After application of the differential-privacy algorithm, the redistricting data will remain “the best population data available” – indeed, Plaintiffs have not pointed to any other extant data that would be better – and, absent a source of better data, the redistricting data will constitute “the only basis for good-faith attempts to achieve population equality.” *Karcher*, 462 U.S. at 738.

Nonetheless, in an effort to show some sort of injury-in-fact, Plaintiffs contend – citing a short law journal article written by a law clerk – that if Alabama were to redistrict based on data subject to the differential-privacy algorithm, “litigation against the State” will be “especially likely.” Compl. ¶ 144. But Plaintiffs do not explain what source of alternative data could undergird such imagined lawsuits. And in all events, “[a]llegations of injury based on predictions regarding future legal proceedings are . . . too speculative to invoke the jurisdiction of an Article III Court.” *Platte River Whooping Crane Critical Habitat Maint. Tr. v. Fed. Energy Regulatory Comm’n*, 962 F.2d 27, 35 (D.C. Cir. 1992). Indeed, the Supreme Court has “been reluctant to endorse standing theories that require guesswork as to how independent decisionmakers will exercise their judgment.” *Clapper v. Amnesty Int’l USA*, 568 U.S. 398, 413 (2013). Moreover, injuries-in-fact must be “real, immediate, and direct.” *Ga. Republican Party v. SEC*, 888 F.3d 1198, 1202 (11th Cir. 2018). And “[a]lthough imminence is concededly a somewhat elastic concept, it cannot be

stretched beyond its purpose, which is to ensure that the alleged injury is not too speculative for Article III purposes – that the injury is *certainly* impending.” *Clapper*, 568 U.S. at 409 (emphasis in original). Alabama’s supposed injury – the possibility of future litigation brought by third parties on a speculative basis at some point in the distant future – cannot support standing.

*Second*, even if Alabama believes that it cannot use the redistricting data as produced by the Secretary, Alabama law does not obligate Alabama to use that data in drawing districts. “While the use of census data is the general practice, no stricture of the federal government requires States to use decennial census data in redistricting, so long as the redistricting complies with the Constitution and the Voting Right Act.” *Ohio v. Raimondo*, No. 3:21-cv-064, 2021 WL 1118049, at \*8 (S.D. Ohio Mar. 24, 2021), *appeal filed*, No. 21-3294 (6th Cir. docketed Mar. 25, 2021); *see Burns v. Richardson*, 384 U.S. 73, 91 (1966) (“[T]he Equal Protection Clause does not require the States to use total population figures derived from the federal census as the standard by which this substantial population equivalency is to be measured.”); *Tucker v. U.S. Dep’t of Commerce*, 958 F.2d 1411, 1418 (7th Cir. 1992) (Posner, J.) (“[S]tates are not required to use census figures for the apportionment of their legislatures.”). Rather, States are required to use “the best population data available” to redistrict, *City of Detroit v. Franklin*, 4 F.3d 1367, 1374 (6th Cir. 1993) – and that data does not necessarily have to derive from the decennial census.

And, in fact, nothing in Alabama’s Constitution requires that the State use U.S. census data for its state legislative apportionment or redistricting. To be sure, Plaintiffs argue that the Alabama Constitution: (i) “requires that the State Legislature use the number of inhabitants, as reported by the Census Bureau, to apportion the seats in the State House and State Senate,” and (ii) obligates “[t]he Legislature [to] conduct legislative redistricting based on the Census Bureau’s tabulations.” Mot. 7 (citing Ala. Const. §§ 197–200). But neither proposition is correct.

*First*, Alabama’s Constitution expressly provides that the State’s apportionment need not necessarily be based on U.S. census data. Though section 198 provides that Alabama’s representatives shall be apportioned among the State’s counties “according to the number of inhabitants in them . . . as ascertained by the decennial census of the United States,” Ala. Const. § 198, section 201 – which Plaintiffs conspicuously neglect to mention – provides in part that if the decennial census is not “full and satisfactory” to the State, then “the legislature shall have the power at its first session after the time shall have elapsed for the taking of said census, to provide for an enumeration of all the inhabitants of this state, upon which it shall be the duty of the legislature to make the apportionment of representatives and senators.” Ala. Const. § 201. Plaintiffs allege that the Alabama Legislature’s “first session after taking the decennial census of the United States’ began February 2, 2021, and will adjourn May 30.” Compl. ¶ 71. And this very lawsuit reflects that in Alabama’s view, the decennial census is not “full and satisfactory” to the State. Accordingly, Alabama’s Legislature *is currently empowered* to conduct its own statewide census, after which “it shall be the duty of the legislature to make the apportionment of representatives and senators.” Ala. Const. § 201.

*Second*, no provision of Alabama’s constitution obligates “[t]he Legislature [to] conduct legislative redistricting based on the Census Bureau’s tabulations.” *Contra* Mot. 7 (citing Ala. Const. §§ 199–200). Sections 199 obligates the legislature to conduct a new apportionment of representatives “after each . . . decennial census.” Ala. Const. § 199. Section 200 obligates the legislature “to divide the state into as many senatorial districts as there are senators” “after each . . . decennial census.” Ala. Const. § 200. Neither section refers to – let alone requires – the *use* of U.S. census data. *See id.*

Simply put, nothing in Alabama’s constitution obligates the State to use census data to fulfill its “sovereign interest in drawing fair districts.” Mot. 33. Rather, if Alabama (incorrectly) believes that the future census redistricting data will be unsuitable for apportionment and redistricting, Alabama may conduct its own census. *See* Ala. Const.

§ 201. And in that case, Alabama’s decision *not* to conduct its own census is a classic “self-inflicted harm” that “does not amount to an ‘injury’ cognizable under Article III.” *Nat’l Family Planning & Reproductive Health Ass’n, Inc. v. Gonzales*, 468 F.3d 826, 831 (D.C. Cir. 2006).

The United States District Court for the Southern District of Ohio recently arrived at a similar conclusion. In *Ohio v. Raimondo*, the State of Ohio sued Defendants, arguing “that the Census Bureau’s plan to deliver redistricting data by September 30, 2021 is contrary to the deadlines established in 13 U.S.C. § 141(c).” *Ohio*, 2021 WL 1118049, at \*6. Like Alabama here, Ohio argued that the September delivery date impeded its sovereign interests. But just like Alabama’s constitution, Ohio’s constitution also “contemplates ways in which redistricting can be accomplished in the absence of census data.” *Id.* Because Ohio’s laws were not actually “frustrated or rendered invalid by the delay in census data,” “[t]he absence of census data thus does not stop the state from implementing its constitutional scheme or otherwise impinge on its sovereign interests in effectuating its law.” *Id.* at \*7. The same analysis applies here.

To be clear, Defendants are not suggesting that Alabama actually conduct its own census. To Defendants’ knowledge, Alabama has no such expertise. But Alabama’s constitution expressly empowers the State to conduct its own census if it is displeased with this year’s decennial census—and if Alabama’s census produces better data than the decennial census, Alabama may use its census to redistrict. Alabama’s concerted decision not to avail itself of its own constitutional powers is a classic self-inflicted injury that cannot support standing.

Pointing to *Karcher v. Daggett*, 462 U.S. 725 (1983), Plaintiffs also suggest—contrary to the Alabama constitution—that the decennial census “‘is the only basis for good-faith attempts to achieve population equality.’” Mot. 33 (quoting *Karcher*, 462 U.S. at 738). But Plaintiffs misread *Karcher*. “The Court in *Karcher* did not hold that the states must use census figures to reapportion congressional representation.” *City of Detroit*, 4 F.3d at

1374. “The Supreme Court merely reiterated a well-established rule of constitutional law: states are required to use the ‘best census data available’ or ‘the best population data available’ in their attempts to effect proportionate political representation.” *Id.* And “[i]f figures other than the census count are the best population data available, the Supreme Court did not, in *Karcher*, bar their use.” *Id.*

**c. Federal Funding**

Plaintiffs allege that “[d]ecennial census data are also used in many federal funding formulas that distribute federal funds to states and localities each year.” Compl. ¶ 148; *see generally id.* ¶¶ 148–158. But Plaintiffs conspicuously do not allege that Alabama is likely—let alone substantially likely—to suffer a *loss* of federal funds based on the application of differential privacy. Indeed, Plaintiffs make no effort to plausibly allege that the level of noise that the differential-privacy algorithm will inject into the future redistricting data will suffice to move the needle on even a *single* source of Alabama’s federal funding—let alone move the needle in a manner that will actually injure the State. Instead, Plaintiffs merely allege (in conclusory fashion) that purported funding variables “will be *affected* by differential privacy” and that such supposed “variance will directly *affect* the amount of federal funding Alabama and its citizens receive.” *Id.* ¶¶ 152, 158 (emphases added). Even assuming these naked allegations could surmount the plausibility threshold, they do not suffice to show substantial risk of *injury*.

In fact, Plaintiffs’ own expert strongly suggests that, to the extent that Alabama’s funding would be affected by differential privacy, it will result in a *windfall* to the State. Plaintiffs allege that “the rural population rate is a primary determinant of where federal spending is allocated.” Compl. ¶ 157. And Plaintiffs’ expert Dr. Barber opines that “[p]laces with fewer people (rural locations) . . . are more likely to be impacted” by the application of differential privacy—and the impact is (in his opinion) that rural areas would *gain* population: that “small [census] blocks, on average, get bigger” and “the largest blocks, on average, get smaller.” Barber Rep., Doc. 3–5, at 13–14; *see also id.* at 15



(quoting the State of Washington: “There is a bias in the demonstration data that causes areas with small populations to get larger while areas with larger populations get smaller.”); *id.* (quoting the State of Utah: “We observe that the population loss in our cities and towns are re-allocated to unincorporated, rural areas of the state.”).

In their motion, Plaintiffs also argue that differential privacy will result in the misallocation of federal funds. *See* Mot. 52–55. But like the challenge to the census rejected by the Supreme Court for lack of standing and ripeness in *Trump v. New York*, 141 S. Ct. 530 (2020), Plaintiffs’ supposed funding “injuries” are also “riddled with contingencies and speculation that impede judicial review.” *Id.* at 535. Plaintiffs’ “misallocation” arguments mirror the arguments improperly accepted by the *New York* district court. *See, e.g., New York v. Trump*, 485 F. Supp. 3d 422, 451 (S.D.N.Y. 2020) (“degraded census data jeopardizes various sovereign interests in allocating funds and administering public works through programs that rely on quality census data”), *vacated and remanded*, 141 S. Ct. 530 (2020). And though the Supreme Court’s dissenters argued that the *New York* plaintiffs’ predictions about the allocation of federal funds should be sufficient for standing purposes, *see* 141 S. Ct. at 540 (Breyer, J., dissenting), the majority rejected that argument. *See id.* at 536 (“The impact on funding is no more certain. According to the Government, federal funds are tied to data derived from the census, but not necessarily to the apportionment counts addressed by the memorandum. . . . Under that view, changes to the Secretary’s § 141(b) report or to the President’s § 2a(a) statement will not inexorably have the direct effect on downstream access to funds or other resources predicted by the dissent.”) (citation omitted).

Just as in *New York*, Plaintiffs’ allegations and arguments regarding a supposed “substantial risk’ of reduced . . . federal resources” “involve[] a significant degree of guesswork.” 141 S. Ct. at 535–36. But the future application of differential privacy, like the future application of the presidential memorandum at issue in *New York*, will not

“predictably change the count.” *Id.* at 536 (emphasis added). Accordingly, Plaintiffs’ “prediction about future injury [is] just that—a prediction.” *Id.*

**d. Vote Dilution**

Plaintiffs also argue that “[t]he Census Bureau’s decision to apply differential privacy . . . creates a substantial risk that” the individual plaintiffs “will have their votes in local, state, and federal elections diluted.” Mot. 36. But “injury results only to those persons domiciled in the under-represented voting districts.” *Wright v. Dougherty Cnty.*, 358 F.3d 1352, 1355 (11th Cir. 2004) (per curiam) (quoting *Fairley v. Patterson*, 493 F.2d 598, 603 (5th Cir. 1974)). Individuals who “have not suffered any harm or injury by the mal-apportioned voting districts” lack standing. *Id.*; see also, e.g., *Common Cause v. Rucho*, 279 F. Supp. 3d 587, 610 n.7 (M.D.N.C. 2018) (three-judge court) (“Plaintiffs in underpopulated districts lack standing to challenge a districting plan on one-person, one-vote grounds.”) (citing *Fairley*, 493 F.2d at 603–04), *vacated and remanded on other grounds*, 138 S. Ct. 2679 (2018).

The individual plaintiffs do not know how the future application of the differential-privacy algorithm will affect the population counts at any level of census geography. Indeed, each of them declares that they do not presently know, “and, in fact, may never know . . . if [their] vote is being weighed as equally as the vote of another voter in a neighboring district.” Williams Decl., Doc. 3–9, ¶ 12; see Green Decl., Doc. 3–10, ¶ 16 (substantially similar); Aderholt Decl., Doc. 3–11, ¶ 26 (substantially similar). At best, Plaintiffs’ argument reduces to the notion that the individual plaintiffs’ votes *may* be diluted. But the Supreme Court’s decisions “are consistent in recognizing a high standard for the risk-of-harm analysis.” *Muransky v. Godiva Chocolatier, Inc.*, 979 F.3d 917, 927 (11th Cir. 2020) (en banc). And “[a]llegations of possible future injury are not sufficient.” *Clapper*, 568 U.S. at 409. See, e.g., *Mont. Env’tl. Info. Ctr. v. Stone-Manning*, 766 F.3d 1184, 1189 n.4 (9th Cir. 2014) (45% chance of harm “does not suffice to show a substantial risk”).

**e. Section 209**

Plaintiffs also assert injury based on the supposed violation of § 209 of the 1998 Appropriations Act, Pub. L. No. 105–119. *See* Mot. 36–38. No such injury exists.

Section 209 provides in part that “[a]ny person aggrieved by the use of any statistical method in violation of the Constitution or any provision of law . . . in connection with the 2000 or any later decennial census, to determine the population for purposes of the apportionment or redistricting of Members in Congress, may in a civil action obtain declaratory, injunctive, and any other appropriate relief against the use of such method.” 1998 Appropriations Act, § 209(b). Even assuming *arguendo* that Plaintiffs constitute such “person[s] aggrieved,” the Eleventh Circuit has made clear that “alleging a statutory violation is not enough to show injury in fact.” *Muransky*, 979 F.3d at 924. And *U.S. House of Representatives* demonstrates this principle in the § 209 context. In that case, the Supreme Court indicated that § 209 “eliminated . . . prudential concerns,” *see* 525 U.S. at 328 – and then proceeded to explain that “the only open justiciability question in this case is whether appellees satisfy the requirements of Article III standing.” *Id.* at 329. If a mere statutory violation of § 209 were sufficient to create Article III standing, the Court’s standing analysis, *see U.S. House of Representatives*, 525 U.S. at 329–34, would have been entirely unnecessary. *See also Muransky*, 979 F.3d at 928 (“A conclusory statement that a statutory violation caused an injury is not enough.”).

**2. Plaintiffs Are Not Injured by Delayed Redistricting Data**

Plaintiffs argue that Alabama is injured by the “delay in producing the population tables.” Mot. 55. “When the federal government prevents a State from applying state law,” they argue, “the State suffers an irreparable harm.” *Id.* (citing *Maryland v. King*, 133 S. Ct. 1, 3 (2012) (Roberts, C.J., in chambers)). But as explained above, Defendants are not preventing Alabama from complying with its own state law, because Alabama’s own constitution does not require census data for redistricting purposes.

Plaintiffs also argue that “delivering redistricting data on September 30 will also likely leave Alabama’s Boards of Registrars at most only four months for reassigning their respective counties’ registered voters to their correct precincts and districts,” yet “[t]he reassignments typically take up to six months.” Mot. 56. But the Boards of Registrars can get started right now with information that the Census Bureau has already provided to Alabama. *See, e.g.,* Whitehorne Decl. ¶¶ 10–12. And Plaintiffs’ declarant also makes clear that the State can “push[] back [its] primary election” by seven weeks. Helms Decl., Doc. 3–3, ¶¶ 14–15. In all events, this is just another way of saying that the 2020 decennial census is not “full and satisfactory” to the State of Alabama, thus empowering Alabama’s legislature to “provide for an enumeration of all the inhabitants” of the State. Ala. Const. § 201. In any event, Plaintiffs – citing the Helms declaration – argue that the Secretary’s September delivery of redistricting data “will result” in one or more harms. Mot. 56 (emphasis added). But the Helms declaration they cite is not so definitive. Rather, the Helms declaration states that “[r]equiring the Boards of Registrars and county commissions to complete the reassignment process on an abbreviated schedule *could* result in one or more” harms. Helms Decl., Doc. 3–3, ¶ 12 (emphasis added). This equivocal declaration cannot support standing: “threatened injury must be *certainly impending* to constitute injury in fact”; “[a]llegations of *possible* future injury are not sufficient.” *Clapper*, 568 U.S. at 409 (emphases in original).

Plaintiffs also argue that “the Bureau’s delay harms” Representative Aderholt “by effectively reducing by at least four months the amount of time [he] can spend campaigning and fundraising.” Mot. 56; *see also* Compl. ¶ 197. But “[t]o establish standing, an injury in fact must be concrete.” *Salcedo v. Hanna*, 936 F.3d 1162, 1167 (11th Cir. 2019) (footnote omitted). In turn, “[a] ‘concrete’ injury must be ‘*de facto*’; that is, it must actually exist.” *Id.* Representative Aderholt’s supposed injury does not meet this standard. Plaintiffs do not contend that these lost months will make it less likely for Representative Aderholt to win reelection. And it is clear why: delayed redistricting data affects *every*

*candidate*—not just Representative Aderholt. In fact, as the incumbent, Representative Aderholt is perhaps likely to *benefit* from a shorter campaign cycle. In all events, Representative Aderholt cannot be said to be injured by the delay in producing redistricting data.

**B. Plaintiffs’ Alleged Injuries Are Not Traceable to Defendants’ Actions**

**1. Plaintiffs’ Alleged Injuries Cannot Be Traced to Defendants’ Plan to Use Differential Privacy**

For similar reasons, Plaintiffs fail to establish the requisite “causal connection between” their alleged injuries and the actions they challenge—*i.e.*, they cannot show that any alleged injury is “fairly . . . trace[able]” to Defendants’ actions. *Lujan v. Defenders of Wildlife*, 504 U.S. 555, 560 (1992). Specifically, Plaintiffs have failed to show that their alleged injuries related to redistricting—*i.e.*, Alabama’s “sovereign interest in drawing fair districts” and the individual plaintiffs’ interest in not having their votes diluted—are traceable to Defendants. *See* Mot. 33, 36. The Supreme Court has explained in no uncertain terms that “[r]edistricting is primarily the duty and responsibility of the State,” *Abbott v. Perez*, 138 S. Ct. 2305, 2324 (2018), and “involves choices about the nature of representation with which [courts] have been shown no constitutionally founded reason to interfere,” *Burns*, 384 U.S. at 92 (emphasis added). “While the use of census data is the general practice, no stricture of the federal government requires States to use decennial census data in redistricting, so long as the redistricting complies with the Constitution and the Voting Rights Act.” *Ohio*, 2021 WL 1118049, at \*8. Thus, in dismissing the State of Ohio’s recent lawsuit against Defendants, Judge Rose concluded that Ohio’s alleged injuries were not traceable to Defendants’ challenged actions, but rather Ohio’s “independent decision to create a state redistricting timeline without the flexibility to accommodate the COVID-19 pandemic.” *Id.*

Here, Alabama’s timetables do not even appear to be incompatible with a September 30, 2021, release of redistricting data. *See* Helms Decl., Doc. 3-3, ¶¶ 14-15 (conceding

that the State can “push[] back” its primary by seven weeks). And in all events, Plaintiffs’ claimed injuries here could only occur if the Alabama legislature declines to exercise its power, in the event that the U.S. decennial census is “not full and satisfactory,” “to provide for an enumeration of all the inhabitants of th[e] state.” Ala. Const. § 201. So any purported injury Alabama may suffer is “fairly . . . trace[able]” to the Alabama legislature’s independent decision to use U.S. census data and the State’s failure to adjust its own timetables, not “the challenged action of the defendant.” *Lujan*, 504 U.S. at 560.

Moreover, even if the Alabama legislature were required to use U.S. census data, Plaintiffs cannot demonstrate traceability because they cannot show that differential privacy will result in data that is less accurate when “compared to a feasible, alternative methodology,” *Nat’l Law Ctr. on Homelessness & Poverty v. Kantor*, 91 F.3d 178, 183 (D.C. Cir. 1996) (emphasis omitted), or that the difference between the two methodologies is sufficiently large to produce some kind of harm, *id.* at 185–86; *see also Franklin v. Massachusetts*, 505 U.S. 788, 802 (1992) (plurality) (challengers to the allocation of overseas employees among states had “neither alleged nor shown . . . that [they would] have had an additional Representative if the allocation had been done using some other source of ‘more accurate’ data” and accordingly did not have standing “to challenge the accuracy of the data used in making that allocation”). As noted above, Plaintiffs maintain that differential privacy will result in inaccurate numbers, but they have identified no other feasible, Census Act-compliant disclosure-avoidance methodology that would produce more accurate numbers. While Plaintiffs note that the Census Bureau has relied on other disclosure-avoidance methods in the past, Mot. 9–12, Dr. Abowd’s declaration explains in detail why those methods are not feasible for the 2020 Census. *See* Abowd Decl. ¶¶ 41–43, 50–51. Absent a feasible alternative, Plaintiffs cannot contend that any alleged inaccuracy is, in fact, “caused” by differential privacy.

## **2. Plaintiffs' Alleged Injuries Cannot Be Traced to Defendants' Delay in Producing Redistricting Data**

Plaintiffs also fail to establish traceability for their purported injuries allegedly arising out of the Bureau's delay in producing redistricting data. Again, because redistricting is ultimately the responsibility of the State, Plaintiffs cannot show that their purported injuries are traceable to the challenged actions of Defendants, as opposed to the State's independent decisions. For this reason, the *Ohio* court recently dismissed Ohio's delay claim on traceability grounds, 2021 WL 1118049, at \*8, and because the same analysis applies here, this Court should do the same.

Plaintiffs also cannot establish traceability because they identify no feasible alternative to producing redistricting data by September 30, 2021. Plaintiffs suggest in passing that the Bureau could have "attempted to deliver apportionment and redistricting numbers to different States 'on a flow basis,'" "prioritizing the States whose laws rely on timely receipt of census data." Mot. 47. But that would place Alabama last in line as its constitution affords the State an alternative path. *See* Ala. Const. § 201; *see generally* Part I.A.1.b. In all events, as the Whitehorne declaration explains, even if the Census Bureau prioritized Alabama's redistricting data to the detriment of the other 49 States, "it would not be able to deliver the data more than a few weeks earlier than a single national release"; "[t]he resulting data may have uncaught errors from [having] been rushed through review without the benefit of review of all States at once"; and it would "delay the release of data for the other 49 states." Whitehorne Decl. ¶¶ 29–30. Because there is no feasible alternative, Plaintiffs cannot contend that their alleged injuries are "caused" by any action by the Bureau.

## **C. Plaintiffs' Purported Injuries Are Not Redressable**

An injury is redressable only if "a decision in a plaintiff's favor would 'significantly increase the likelihood' that [plaintiff] would obtain relief that directly redresses the injury that [plaintiff] claims to have suffered." *Lewis v. Governor of Ala.*, 944 F.3d 1287,

1301 (11th Cir. 2019) (en banc). Plaintiffs must demonstrate not only that they have suffered an injury that is traceable to Defendants, but also that “redress is likely ‘as a practical matter.’” *Jacobson v. Fla. Sec’y of State*, 974 F.3d 1236, 1255 (11th Cir. 2020) (quoting *Utah v. Evans*, 536 U.S. 452, 461 (2002)). Here, Plaintiffs cannot demonstrate that any of their alleged injuries would be redressed by an order enjoining Defendants from using differential privacy or requiring Defendants to produce redistricting data sooner than is possible.

**1. Enjoining Differential Privacy Would Not Redress Plaintiffs’ Alleged Injuries**

An order enjoining the Census Bureau from using differential privacy for the 2020 Census would not “significantly increase the likelihood” that Plaintiffs’ alleged injuries would be redressed. To the contrary, there is a significant likelihood that an order enjoining differential privacy would only make any alleged injuries worse. If the Court were to enjoin differential privacy, the Census Bureau would still need to comply with sections 8 and 9 of the Census Act, which prohibit Defendants from “disclos[ing] the information reported by, or on behalf of, any particular respondent,” or “mak[ing] any publication whereby the data furnished by any particular establishment or individual . . . can be identified.” 13 U.S.C. §§ 8(b), 9(a)(2). But the Census Bureau cannot rely solely on the disclosure avoidance methods used in the 2010 Census, which would also allow individual respondents’ data to be identified. *See* Abowd Decl. ¶¶ 38–39.

To comply with sections 8 and 9 of the Census Act, the Census Bureau would instead have to “swap” or “suppress” data at the census block level. *Id.* ¶¶ 40–43. This would *exacerbate* Plaintiffs’ alleged injuries, not redress them, because “[b]oth choices would delay results and diminish accuracy.” *Id.* ¶ 84. For example, Plaintiffs allege that differential privacy “impede[s] the State’s sovereign interest in drawing fair districts.” Mot. 33. As explained above, differential privacy will not cause any such injury to Alabama’s sovereign interests. *See supra*, Part I.A.1.b. By contrast, swapping or suppression



at the levels necessary to protect the census data could very well impede Alabama's ability to draw fair districts. *See* Abowd Decl. ¶¶ 42, 43, 87. Thus, "as a practical matter," an order enjoining differential privacy is not likely to redress Plaintiffs' claimed injuries resulting from allegedly inaccurate data. *Jacobson*, 974 F.3d at 1255.

An order enjoining the use of differential privacy would also only extend the Bureau's delay in providing redistricting data. As Dr. Abowd explains, it would take the Bureau "multiple months" to develop, test, and implement any alternative disclosure-avoidance methodology. Abowd Decl. ¶ 85. Accordingly, the relief that Plaintiffs seek—an order enjoining differential privacy—would hinder, rather than help, the Bureau's ability to produce redistricting data to the States as soon as possible.

## **2. Requiring the Census Bureau to Produce Redistricting Data Sooner Would Not Redress Plaintiffs' Alleged Injuries**

Nor can Plaintiffs demonstrate redressability as to their delay claim. As Judge Rose observed in holding that the State of Ohio had not demonstrated redressability in its similar challenge to the Census Bureau's delay, "a judicial decree is only the means to an end: 'At the end of the rainbow lies not a judgment, but some action (or cessation of action) by the defendant that the judgment produces.'" *Ohio*, 2021 WL 1118049, at \*5 (quoting *Doe v. DeWine*, 910 F.3d 842, 850 (6th Cir. 2018)). "In other words, '[r]edress is sought *through* the court, but *from* the defendant,' and '[t]he real value of the judicial pronouncement—what makes it a proper judicial resolution of a case or controversy rather than an advisory opinion—is in the settling of some dispute which affects the behavior of the defendant towards the plaintiff.'" *Id.* (quoting *Doe*, 910 F.3d at 850) (emphasis added).

Here, as in *Ohio*, "[Alabama] seeks an advisory opinion that cannot redress their claimed injury." *Id.*; *see also Jacobson*, 974 F.3d at 1255 (redress must be likely "as a practical matter"); *Brown v. Berhndt*, 12-cv-24-KGB, 2013 WL 1497784, at \*5 (E.D. Ark. Apr. 10, 2013) (no standing where "injunctive relief [wa]s impossible"). That's because it is

“not possible under any scenario for the Census Bureau to produce these data at this time or at any time in the immediate future, and the Census Bureau would be unable to comply with any such order from the Court.” Whitehorne Decl. ¶ 14. “[T]he Census Bureau must complete a series of interim steps prior to delivering the redistricting data,” and “[e]ach of these interim steps, in order, is required to move to the next.” *Id.* ¶¶ 15–16. Those steps will likely not be completed until September 30, 2021, though the Bureau expects to be able to make a “legacy” format of the redistricting data file available to States in mid-to-late August. *Id.* ¶¶ 14–16, 27–28. Although the 2020 Census Operational Plan provided for only three months from the planned release of apportionment data on December 31, 2020, *see* Mot. 28, 49, the Bureau now requires five months because of operational changes that the Bureau made to expedite the release of the constitutionally required apportionment counts, including “decoupling” certain processes that the Bureau would have normally completed at the same time. Thieme Decl. ¶¶ 84–86.

Alabama’s purported injury is “also unredressable when it comes to redistricting for congressional (as opposed to state) elections.” *Ohio*, 2021 WL 1118049, at \*5. In order to draw congressional districts, Alabama must first know the number of Representatives it will have in Congress to know how many districts to draw. 2 U.S.C. § 2c. But the Census Bureau has not yet finished, and neither the Secretary nor the President have yet reported, the apportionment of Representatives. Once the President reports the appointment numbers to Congress, apportionment will be entirely in Congress’s hands to accept or reject. *See* 2 U.S.C. § 2a(b) (commanding that apportionment only occurs “under [2 U.S.C. § 2a] or subsequent statute”). So even if the Court ordered the Census Bureau to produce redistricting data immediately, Alabama would be no closer to drawing congressional districts until Congress has determined the number of Representatives to which Alabama is entitled. In such circumstances, redressability (and standing) are lacking. *See Leifert v. Strach*, 404 F. Supp. 3d 973, 982 (M.D.N.C. 2019) (no redressability where

“[i]t is not merely speculative, but rather impossible, for the requested relief to remedy the alleged injury”).

Put simply, Alabama seeks the impossible. But “a court may not require an agency to render performance that is impossible.” *Am. Hosp. Ass’n v. Price*, 867 F.3d 160, 167 (D.C. Cir. 2017). Indeed, “[i]t has long been settled that a federal court has no authority . . . to declare principles or rules of law which cannot affect the matter in issue in the case before it.” *Church of Scientology of Cal. v. United States*, 506 U.S. 9, 12 (1992). The Court should therefore reject Alabama’s request for an advisory opinion based on the hypothetical world in which it were possible for the Census Bureau to comply with Alabama’s requested relief. The Court cannot “order a party to jump higher, run faster, or lift more than she is physically capable.” *Am. Hosp. Ass’n*, 867 F.3d at 168; Whitehorne Decl. ¶ 14 (explaining that “it would be a physical impossibility” to provide redistricting data at this time).

## **II. PLAINTIFFS ARE NOT ENTITLED TO A PRELIMINARY INJUNCTION.**

“A preliminary injunction is an extraordinary remedy never awarded as of right.” *Winter v. Nat. Res. Def. Council, Inc.*, 555 U.S. 7, 24 (2008). Its “chief function . . . is to preserve the status quo until the merits of the controversy can be fully and fairly adjudicated.” *Ne. Fla. Chapter of Ass’n of Gen. Contractors of Am. v. City of Jacksonville*, 896 F.2d 1283, 1284 (11th Cir. 1990). But Plaintiffs are not asking the Court to preserve the status quo. Entering Plaintiffs’ proposed injunction would *upend* the status quo and would effectively constitute final relief in Plaintiffs’ favor by forcing the Census Bureau to completely overhaul its existing disclosure-avoidance methodology and to make wholesale, untested operational changes to produce redistricting data as quickly as possible.

Even assuming that Plaintiffs’ proposed relief could be characterized as a preliminary injunction, Plaintiffs do not satisfy any of the preliminary-injunction standards. “In order to obtain [a preliminary injunction], a party must establish four separate require-

ments – namely, that (1) it has a substantial likelihood of success on the merits; (2) irreparable injury will be suffered unless the injunction issues; (3) the threatened injury to the movant outweighs whatever damage the proposed injunction may cause the opposing party; and (4) if issued, the injunction would not be adverse to the public interest.” *Swain v. Junior*, 961 F.3d 1276, 1284–85 (11th Cir. 2020). And the latter two factors “merge when, as here, the Government is the opposing party.” *Id.* at 1293.

Plaintiffs “bear[] the burden of persuasion to clearly establish all . . . of these prerequisites.” *Wreal, LLC v. Amazon.com, Inc.*, 840 F.3d 1244, 1247 (11th Cir. 2016). “[F]ailure to meet even one dooms” Plaintiffs’ bid for a preliminary injunction. *Id.* at 1248.

**A. Plaintiffs Are Unlikely to Succeed on the Merits of Their Differential Privacy Claims.**

**1. Plaintiffs’ Census Act Claim Is Not Likely to Succeed**

Plaintiffs are not likely to prevail on their § 141(c) claim. *See* Compl. ¶¶ 198–202. As explained above, Defendants’ use of differential privacy will comply with § 141(c). *See supra* Part I.A.1.a.

Moreover, Alabama lacks a private right of action to assert a claim under § 141(c). “Like substantive federal law itself, private rights of action to enforce federal law must be created by Congress.” *Alexander v. Sandoval*, 532 U.S. 275, 286 (2001). “Where Congress has not created a private right of action, courts may not do so, ‘no matter how desirable that might be as a policy matter, or how compatible with the statute.’” *Bellitto v. Snipes*, 935 F.3d 1192, 1202 (11th Cir. 2019) (quoting *Sandoval*, 532 U.S. at 287).

The only private right of action to enforce § 141(c) flows through § 209(b) of the 1998 Appropriations Act.<sup>4</sup> Section 209(b) provides a private right of action to “[a]ny person aggrieved by the use of any statistical method in violation of the Constitution or any

---

<sup>4</sup> In their motion, Plaintiffs seem to suggest that § 209(b) provides them with a separate substantive claim. *See, e.g.*, Mot. 37–38 (“Defendants have violated Plaintiffs’

provision of law . . . in connection with the 2000 or any later decennial census, to determine the population for purposes of the apportionment or redistricting of Members in Congress.” Even assuming *arguendo* that differential privacy constitutes a “statistical method” as defined in § 209, Alabama is not a “person aggrieved.”

Section 209 states that “an aggrieved person . . . includes – (1) any resident of a State whose congressional representation or district could be changed as a result of the use of a statistical method challenged in the civil action; (2) any Representative or Senator in Congress; and (3) either House of Congress.” 1998 Appropriations Act § 209(d). Absent from this list of “aggrieved person[s]” are “States.” Plaintiffs nonetheless argue that the Court should infer that “Alabama is an ‘aggrieved person,’ too.” Mot. 37. But Congress did not include “States” in its list of “aggrieved persons,” and for this Court to do so would run counter to the “longstanding interpretive presumption that ‘person’ does not include the sovereign.” *Return Mail, Inc. v. U.S. Postal Serv.*, 139 S. Ct. 1853, 1861–62 (2019). For this reason, there is a “background presumption that States are not ‘persons.’” *Cook Cnty. v. United States ex rel. Chandler*, 538 U.S. 119, 133 n.10 (2003); see *Vt. Agency of Nat Res. v. United States ex rel. Stevens*, 529 U.S. 765, 780–88 (2000) (State is not a “person” for False Claims Act purposes). And “although the presumption is not a hard and fast rule of exclusion . . . it may be disregarded only upon some affirmative showing of statutory intent to the contrary.” *Return Mail, Inc.*, 139 S. Ct. at 1862.

If anything, the statutory text reflects Congress’s intent to *exclude* States from the definition of aggrieved persons. After all, this is not a situation where Congress left the term “person” undefined. Rather, Congress enacted a specific definition of “aggrieved

---

rights under Public Law No. 105–119, § 209(b).”). But Plaintiffs do not assert a claim for violation of § 209(b). See generally Compl. ¶¶ 198–241. And for good reason: Section 209(b) simply creates a private right of action. See *Common Cause v. Trump*, No. 1:20-cv-02023, -- F. Supp. 3d --, 2020 WL 8839889, at \*12 (D.D.C. Nov. 25, 2020) (three-judge court); *Glavin v. Clinton*, 19 F. Supp. 2d 543, 547 (E.D. Va. 1998) (three-judge court), *aff’d sub nom.*, *Dep’t of Commerce v. U.S. House of Representatives*, 525 U.S. 316 (1999).

person” in § 209(d). That definition even included “either House of Congress” – hardly within the usual definition of “person.” But despite the Supreme Court’s “background presumption that States are not ‘persons,’” *Cook Cnty.*, 538 U.S. at 133 n.10, Congress – which is presumed to “legislate[] with knowledge of [the Supreme Court’s] basic rules of statutory construction,” *McNary v. Haitian Refugee Ctr., Inc.*, 498 U.S. 479, 496 (1991) – declined to include “States” in its definition of “aggrieved person.”

Plaintiffs acknowledge that States are “not expressly named in the statute,” but nonetheless have argued that “[t]he statute’s natural reading includes the States alongside Section 209(d)’s enumerated parties.” Pls. Mot., Doc. 2, at 5–7. Hardly. Given (i) the background presumption that “persons” do not include States, and (ii) Congress expressly included its Houses in defining “aggrieved person[s]” yet did not “expressly” include States, the “statute’s natural reading” is that “aggrieved person[s]” do not include “States.” Plaintiffs also argue that “a contrary interpretation would contravene the statute’s purpose.” Pls. Mot., Doc. 2, at 6. Even assuming Plaintiffs could be considered the arbiters of congressional purpose, “it is ultimately the provisions of our laws rather than the principal concerns of our legislators by which we are governed.” *Oncale v. Sundowner Offshore Servs., Inc.*, 523 U.S. 75, 79 (1998).

Nor can Plaintiffs rely on the fact that the “aggrieved person” is defined as “includ[ing]” various persons and entities. 1998 Appropriations Act § 209(d). After all, the Dictionary Act defines “person” as “includ[ing] corporations, companies, associations, firms, partnerships, societies, and joint stock companies, as well as individuals,” 1 U.S.C. § 1 (emphasis added) – yet the Supreme Court held that “[t]he absence of any comparable provision extending the term to sovereign governments implies that Congress did not desire the term to extend to them.” *United States v. United Mine Workers of Am.*, 330 U.S. 258, 275 (1947).

In sum, Alabama cannot take advantage of § 209's narrow right of action to enforce § 141(c), and in any event, none of the Plaintiffs are likely to succeed on their § 141(c) claims. *See supra* Part I.A.1.a.

**2. The Individual Plaintiffs' Equal Protection Claim Is Not Likely to Succeed**

The individual plaintiffs are not likely to succeed on their one-person-one-vote equal-protection claim. *See* Mot. 35–36. Only individuals residing in under-represented voting districts may bring one-person-one-vote claims. *Wright*, 358 F.3d at 1355. And “over-represented voting district members are barred from bringing suit on behalf of persons who reside in under-represented voting districts.” *Id.* Even assuming *arguendo* that census operational decisions could be susceptible to vote-dilution challenges, Plaintiffs have made clear that they do not know — “and, in fact, may never know” — whether their votes will be diluted. Williams Decl., Doc. 3–9, ¶ 12; Green Decl., Doc. 3–10, ¶ 16; Aderholt Decl., Doc. 3–11, ¶ 26. Plaintiffs concede that they cannot demonstrate any actual or impending vote dilution, and are thus unlikely to succeed on their vote-dilution claims.

Plaintiffs also have not pointed the Court to any case where census operations were enjoined on the grounds that resulting census data might lead States to redistrict in a manner that violated the one-person-one-vote principle. And, in fact, the Supreme Court has rejected such a bid. *See Wisconsin v. New York*, 517 U.S. 1, 16–17 (1996) (“[T]he ‘good-faith effort to achieve population equality’ required of a State conducting intrastate redistricting does not translate into a requirement that the Federal Government conduct a census that is as accurate as possible.”). This is not surprising. As explained above, “the Equal Protection Clause does not require the States to use total population figures derived from the federal census as the standard by which this substantial population equivalency is to be measured.” *Burns*, 384 U.S. at 91. Indeed, Alabama’s own constitution empowers the State to conduct its own census if it is dissatisfied with the decennial census. Ala. Const. § 201. So to the extent that the application of differential privacy

could be said to cause any “vote dilution,” the decision to use federal census data is Alabama’s alone, and no equal-protection claim may lie against the Defendants.

**3. Plaintiffs’ APA Challenges to Differential Privacy Are Not Likely To Succeed**

Plaintiffs’ APA claims face a fundamental problem: the Census Bureau has not yet finalized critical details on how it will use differential privacy. Plaintiffs acknowledge this. *See, e.g.*, Mot. 1 (describing differential privacy as a “still developing confidential algorithm”); Bryan Rep., Doc. 3–6, at 7 (claiming that “[t]he Census Bureau . . . will make a final decision about how DP will be implemented in the redistricting data by early May 2021”). The “in-progress” nature of differential privacy dooms Plaintiffs’ APA claim because this Court lacks jurisdiction when there is no final agency action. *See Nat’l Parks Conservation Ass’n v. Norton*, 324 F.3d 1229, 1236 (11th Cir. 2003).

Plaintiffs try to get around this problem by styling their legal theory as a facial challenge to differential privacy, basing their claim on the 2018 Operational Plan that announced the Census Bureau intended to use differential privacy but that left the critical details to be filled in later. *See* Mot. 40. But the core of Plaintiffs’ concerns relate to the Census Bureau’s later and still ongoing choices like setting the specific privacy-loss budget. And in any event, even if Plaintiffs’ claims (APA or otherwise) were proper and could be characterized as a facial challenge to the 2018 Operational Plan, they would run headlong into the doctrine of laches. *See infra* Part II.A.4.

**a. The Differential Privacy Announcement Was Not Final Agency Action**

No “agency action” as defined by the APA. A cognizable APA claim must challenge a “circumscribed, discrete agency action[]” and it cannot advance a “broad programmatic attack” on an agency’s operations. *Norton v. S. Utah Wilderness All.*, 542 U.S. 55, 61–62 (2004) (“SUWA”); *see also* 5 U.S.C. § 551; 5 U.S.C. § 701(b)(2) (agency action includes “an agency rule, order, license, sanction, relief, or the equivalent or denial thereof”). Put differently, the APA does not permit a plaintiff to attack an agency program “consisting



of . . . many individual actions” simply by characterizing it as “agency action” under the APA. *Lujan v. Nat’l Wildlife Fed’n*, 497 U.S. 871, 893 (1990). While “[c]ourts are well-suited to reviewing specific agency decisions,” they are “woefully ill-suited [ ] to adjudicate generalized grievances asking [them] to improve an agency’s performance or operations.” *City of New York v. U.S. Dep’t of Def.*, 913 F.3d 423, 431 (4th Cir. 2019).

The Census’s data-processing operations, including disclosure avoidance, “expressly are tied to one another,” so altering any of these operations “would impact the efficacy of the others, and inevitably would lead to court involvement in ‘hands-on’ management of the Census Bureau’s operations.” *NAACP v. Bureau of the Census*, 945 F.3d 183, 191 (4th Cir. 2019) (citing *SUIWA*, 542 U.S. at 66–67), *aff’g in part and rev’g in part*, 399 F. Supp. 3d 406 (D. Md. 2019); *see, e.g.*, Whitehorne Decl. ¶¶ 15–16, 21; Abowd Decl. ¶¶ 84–89. In *NAACP*, plaintiffs challenged certain “design choices” within the Census Bureau’s December 2018 Operational Plan—the same Plan that Plaintiffs here claim was the “final agency action” by announcing that the Bureau intended to use differential privacy. *Compare NAACP*, 945 F.3d at 187–88 n.1 *with* Compl. ¶ 79 n.6. The *NAACP* district court found that the design choices within the Operational Plan were not agency action, explaining that “if the Court were to interject itself into the Bureau’s process during the critical final preparations, requiring—as Plaintiffs request—its monitoring and approval of the plans along the way, it is hard to imagine that this oversight would not hinder the process as opposed to facilitate it.” *NAACP v. Bureau of the Census*, 382 F. Supp. 3d 349, 372 (D. Md. 2019).

Plaintiffs’ differential privacy challenge fails this same threshold agency-action inquiry because it is a “broad programmatic attack” on the Census Bureau’s disclosure avoidance operations, not a challenge to “circumscribed, discrete agency action[.]” *SUIWA*, 542 U.S. at 61–62. While Plaintiffs style their legal theory as a facial challenge to differential privacy, a close read of their complaint, motion, and expert reports shows

they ask the Court to scrutinize highly technical policy decisions related to how the Census Bureau *might* implement differential privacy. For example, Plaintiffs take issue with what data will remain untouched during the disclosure-avoidance operations—data sets known as “invariants.” Compl. ¶ 89; Mot. 14. They complain that the planned 2020 invariants include “(1) the total population of each State, (2) the total housing units at the census block level, and (3) the number of group quarters facilities by type at the census block level.” Mot. 14 & n.30 (citing a February 2021 summary file). But the 2020 invariants were not finalized in the 2018 Operational Plan and thus are beyond the scope of Plaintiffs’ current APA claims.

The Census Bureau’s policy choices for what data to hold constant when applying differential privacy could have dominoing impacts on both the disclosure avoidance process and the interrelated data-processing steps. *See* Abowd Decl. ¶ 88. So any Court order commanding the Bureau to set particular invariants—or an order changing to a different disclosure-avoidance method altogether—would require “a sweeping overhaul to the [processing operations], which exceeds the scope of reviewable ‘agency action.’” *NAACP*, 399 F. Supp. at 422. Plaintiffs’ requested relief shows the challenged action is not the type of circumscribed agency action that the APA makes reviewable.

*No jurisdiction because no final agency action.* Even if the 2018 decision to use differential privacy constitutes agency action, this Court still lacks jurisdiction over Plaintiffs’ APA claims because that decision was not *final* agency action. *See In re MDL-1824 Tri-State Water Rights Litig.*, 644 F.3d 1160, 1181, 1185 (11th Cir. 2011). To demonstrate subject-matter jurisdiction, Plaintiffs must show that “the administrative action in question is [] ‘final’ within the meaning of 5 U.S.C. § 704.” *Nat’l Parks Conservation Ass’n v. Norton*, 324 F.3d at 1236. To be final agency action, the challenged action must “mark the ‘consummation’ of the agency’s decision-making process—it must not be of a merely tentative or interlocutory nature” and the challenged action “must be one by which rights or obligations have been determined, or from which legal consequences will flow.” *Bennett v.*

*Spear*, 520 U.S. 154, 177–78 (1997); *Tri-State Water Rights*, 644 F.3d at 1181. Plaintiffs fail on both counts.

First, the Supreme Court has held that interim decisions about Census data processing are not complete until the final decision-maker delivers the data. In *Franklin*, Massachusetts challenged a particular method to assign home states for military personnel stationed abroad. *Franklin*, 505 U.S. at 790. The Supreme Court rejected Massachusetts’ challenge, explaining that there was no final agency action until the President delivered the final apportionment count to Congress pursuant to Section 141(b). 505 U.S. at 800. The interim steps taken by the Secretary of Commerce and the Census Bureau prior to the delivery of the final apportionment numbers under § 141(b) were tentative and not final agency action. *Id.*; *see id.* at 799 (“The President, not the Secretary, takes the final action that affects the States.”). The same analysis applies to the redistricting under § 141(c); the interim steps taken by the Census Bureau before the Secretary delivers the redistricting data to the states cannot constitute final action. *See City of Detroit*, 4 F.3d at 1377 n.6. Final action will occur only when the Secretary delivers the final data to the States, which has not yet occurred. Plaintiffs’ contrary position—that the Census Bureau’s operational plan can somehow bind the Secretary of Commerce—has no merit. “There is no authority for the proposition that a lower component of a government agency may bind the decision making of the highest level.” *Cnty. Care Found. v. Thompson*, 318 F.3d 219, 227 (D.C. Cir. 2003).

Even setting aside *Franklin*, the factual issues that Plaintiffs flag in their motion and declarations underscore why there is no final agency action. Plaintiffs and their declarants flag potential issues in non-final, *demonstration* data products—not the final redistricting data. *See generally* Mot. 20–24; Bryan Rep., Doc. 3–6. The entire point of releasing the demonstration products was to identify issues like the ones flagged by Plaintiffs. *See* Abowd Decl. ¶¶ 58–61. Census Bureau officials have explained that they are still working to resolve issues like those identified in the motion and declarations. *See*

*id.* ¶¶ 68–71. In these circumstances where the agency is actively working to resolve known issues, this court should follow the instruction of the Eleventh Circuit, “exercise restraint,” and let the Census Bureau use “its own institutional expertise” to address potential issues before releasing its final product. *LabMD, Inc. v. FTC*, 776 F.3d 1275, 1278 (11th Cir. 2015) (no final agency action when “agency proceeding is ongoing”).

Critical details of how the Census Bureau will implement differential privacy have not yet been finalized. In particular, the privacy-loss budget will not be set until June. Abowd Decl. ¶ 71. Plaintiffs acknowledge that the eventual privacy-loss budget will affect the ultimate redistricting data: “Dialing the [privacy-loss budget] up to infinity results in perfect accuracy but theoretically imperfect privacy, whereas setting the [privacy-loss budget] at zero results in perfect privacy but useless data.” Mot. 13. And Plaintiffs recognize that the Census Bureau has not reached a final decision on this critical matter. See Mot. 40 (“To be sure, the Bureau has yet to set the privacy loss budget it will use—that decision is still in the works.”) (emphasis added); *id.* at 1 (“the Bureau intends to provide numbers produced by a *still developing* confidential algorithm”) (emphasis added); *id.* at 17 (the Bureau “seeks to impose a *still-developing theory of privacy* onto the decennial census”) (emphasis added). Plaintiffs’ expert, Mr. Bryan, was even more blunt: “The Census Bureau . . . will make a *final decision* about how DP will be implemented in the redistricting data by early May 2021.” Bryan Rep., Doc. 3–6, at 7 (emphasis added). The 2018 Operational Plan was not the consummation of decision-making; in many ways, it was just the beginning of a iterative process that is still in progress.

*Second*, even if the 2018 Operational Plan could somehow be considered the consummation of an agency’s decision-making, it is still not “final” under the APA because it does not “determine any rights or obligations and imposes no legal consequences.” *Clayton Cnty. v. FAA*, 887 F.3d 1262, 1266–67 (11th Cir. 2018). The Operational Plan’s announcement that the Census Bureau would use differential privacy was “purely infor-

mational,” “[c]ompell[ed] no one to do anything,” and “had no binding effect whatsoever – not on the agency and not on” the general public. See *Indep. Equip. Dealers Ass’n v. EPA*, 372 F.3d 420, 427 (D.C. Cir. 2004).

The decision to use differential privacy, standing alone, does not cause the purported “legal consequences” claimed by Plaintiffs. Citing no case law, Plaintiffs claim that the 2018 decision to use differential privacy causes legal consequences by supposedly impeding Alabama’s ability to redistrict and creating a “substantial risk” that individual plaintiffs’ constitutional rights will be abridged. Mot. 40. But those purported “legal consequences” do not inherently flow from the use of differential privacy; those purported consequences flow from third-party decisions regarding redistricting – such as Alabama’s decision not to conduct the census for which its own constitution allows. And even if legal consequences flow from the final redistricting data, that final product will depend on the Census Bureau’s ultimate methodology and privacy-loss budget – not the 2018 decision to use differential privacy.

Nor do the supposed accuracy issues flagged by Plaintiffs somehow demonstrate that the decision to use differential privacy had legal consequences. Plaintiffs’ analysis was based on preliminary demonstration data. As Plaintiffs acknowledge, “the Bureau has stated that it intends to set a less conservative privacy loss budget for the final tabulations of population than it did for the demonstration products.” Mot. 35. And thus the final redistricting “numbers will be less skewed than they are in the demonstration data.” *Id.* Until the Census Bureau sets the final privacy-loss budget and releases the final numbers, Plaintiffs have not shown that there will be *any* legal consequences from differential privacy. The mere announcement that the Census Bureau would use differential privacy lacks legal consequence and is not reviewable final agency action under the APA.

**b. Even Assuming the Differential Privacy Announcement Constituted Final Agency Action, It Did Not Violate the APA**

Plaintiffs argue that “[t]he Census Bureau’s decision to adopt differential privacy is contrary to law, contrary to constitutional right, and in excess of statutory authority.” Mot. 40. They premise this argument on the notion that “the *application* of differential privacy to the population tabulations given to the States” is somehow inconsistent with 13 U.S.C. § 141(c) or that it would supposedly “create a substantial risk that individual Plaintiffs will have their equal protection rights violated.” Mot. 40 (emphasis added).

But Plaintiffs cannot challenge the eventual *application* of differential privacy through an APA challenge to the decision to ultimately implement *some form* of differential privacy. Indeed, Plaintiffs’ § 141(c) and equal-protection challenges are premised on the notion that the Census Bureau’s eventual application of differential privacy will not hold sub-state population counts invariant. But, as explained above, the invariants were not finalized in the 2018 Operational Plan and thus are beyond the scope of Plaintiffs’ current APA challenges to the 2018 Operational Plan. And even assuming *arguendo* that the 2018 Operational Plan had finalized invariants for the eventual application of differential privacy, Plaintiffs’ facial APA challenge to that supposed decision still would fail, as Plaintiffs are not likely to succeed on their § 141(c) or equal-protection claims. *See generally supra* Parts I.A.1.a, I.A.1.d, II.A.1, II.A.2.

For similar reasons, Plaintiffs cannot demonstrate that the *decision* to adopt differential privacy is arbitrary and capricious. Plaintiffs hinge their arbitrary-and-capricious APA claim on the notion that the application of differential privacy will supposedly preclude the Secretary from meeting her obligations “to report accurate tabulations of population under subsection 141(c),” Mot. 42—that is, Plaintiffs’ complaint is again about invariants, and not the disclosure-avoidance methodology in the abstract. And as the 2018 Operational Plan did not declare that sub-state population counts would be made variant, any such decision cannot be challenged in Plaintiffs’ APA claim.

And in all events where (unlike here) there is final agency action, the arbitrary and capricious standard is “exceedingly deferential.” *Sierra Club v. Van Antwerp*, 526 F.3d 1353, 1360 (11th Cir. 2008). The Court is “not authorized to substitute [its] judgment for the agency’s as long as its conclusions are rational.” *Miccosukee Tribe of Indians of Fla. v. United States*, 566 F.3d 1257, 1264 (11th Cir. 2009). “A court simply ensures that the agency has acted within a zone of reasonableness and, in particular, has reasonably considered the relevant issues and reasonably explained the decision.” *FCC v. Prometheus Radio Project*, 141 S. Ct. 1150, 1158 (2021). And the Eleventh Circuit “believe[s] it appropriate to give an extreme degree of deference to the agency when it is evaluating scientific data within its technical expertise.” *Nat’l Mining Ass’n v. Sec’y, U.S. Dep’t of Labor*, 812 F.3d 843, 866 (11th Cir. 2016).

As explained *supra*, Background Parts C & D, the Census Bureau determined that the disclosure-avoidance methodologies it previously used to protect census data were no longer sufficient given the rise in computing power, and that differential privacy was “[t]he best disclosure avoidance option that offers a solution capable of addressing the new risks of reconstruction-abetted re-identification attacks, while preserving the fitness-for-use of the resulting data for the important governmental and societal uses of census data.” Abowd Decl. ¶ 47. The Census Bureau’s decision-making process is not arbitrary or capricious.

Plaintiffs’ arbitrary-and-capricious claim is premised on a number of false notions. For starters, Plaintiffs argue that “the Bureau has not shown that traditional disclosure avoidance methods like data swapping are insufficient to meet” the Census Act’s confidentiality requirements. Mot. 41–42. But that position is easily rebutted by the JASON report that Plaintiffs repeatedly cited in their opening brief. *E.g.*, JASON, *Formal Privacy Methods for the 2020 Census* (Apr. 2020) at 6, available [here](#) (“Approaches to disclosure avoidance such as swapping and top and bottom coding applied at the level used in the

2010 census are insufficient to prevent re-identification given the ability to perform database reconstruction and the availability of external data.”); *accord, e.g.*, Abowd Decl. ¶¶ 38–39, 41–43, 50.

Plaintiffs further argue that the Census Bureau “misinterpreted the confidentiality requirements of § 9,” contending that “[c]onfidentiality is only implicated—in theory—when a recipient of census data uses the information published by the Bureau *together with* other datasets” to re-identify respondents. Mot. 43 (emphasis in original). But it is Plaintiffs that misconstrue the Census Act’s confidentiality requirements. Initially, Plaintiffs’ argument fails the plain text of the statute. Section 9(a) provides that Bureau staff, among others, generally may not “make any publication whereby the data furnished by any particular establishment or individual under this title can be identified.” 13 U.S.C. §§ 9(a), (a)(2). And the Census Bureau demonstrated, as corroborated by JASON, that the 2010 disclosure-avoidance methodology resulted—given recent advances in computing power—in publications that allowed respondent data to be identified. Indeed, under Plaintiffs’ atextual reading of § 9, the Census Bureau need not apply *any* disclosure-avoidance mechanism at all—not even to protect the sole, easily-identifiable Filipino American in the 20-person census block in the data-swapping example they provide, *see* Mot. 10–11—because, in their view, the Census Bureau would only violate § 9 if the *Bureau* publishes respondents’ names and addresses.

In all events, Plaintiffs conspicuously ignore § 9’s companion, 13 U.S.C. § 8, as well as on-point Supreme Court precedent. In *Baldrige v. Shapiro*, 455 U.S. 345 (1982), the Supreme Court expressly rejected the argument that the Census Act’s “confidentiality provisions protect raw data only if the individual respondent can be identified.” *Id.* at 355. Rather, “Congress plainly contemplated that raw data reported by or on behalf of individuals was to be held confidential and not available for disclosure.” *Id.*; *see also id.* at 361 (“§ 8(b) and § 9(a) of the Census Act embody explicit congressional intent to preclude *all*



disclosure of raw census data reported by or on behalf of individuals”) (emphasis in original). So while re-identification may not be possible without the use of other sources of data, the Census Bureau’s database-reconstruction experiment demonstrated that its 2010 census publications could be reverse-engineered, and thus resulted in an unfortunate “disclosure of raw census data reported by or on behalf of individuals.” *Id.* at 361.

Nor did Defendants ignore their end-users’ reliance interests. The 2018 Operational Plan itself made clear that the application of differential-privacy constitutes “a delicate balancing act”: “enough noise must be added to protect confidentiality, but too much noise could damage the statistic’s fitness-for-use.” 2018 Operational Plan, Doc. 3–4, at 140. “The Census Bureau decided that differential privacy was the best tool after analyzing the various options through the lens of economics.” *Abowd Decl.* ¶ 41. “Efficiently protecting privacy can be viewed as an economic problem because it involves the allocation of a scarce re-source – confidential information – between two competing uses: public data products and privacy protection.” *Id.* The Bureau’s “empirical analysis showed that differential privacy offered the most efficient trade-off between privacy and accuracy – our calculations showed that the efficiency of differential privacy dominated traditional methods.” *Id.* “In other words, regardless of the level of desired confidentiality, differential privacy will always produce more accurate data than the alternative traditional methods considered by the Census Bureau.” *Id.*

The ultimate accuracy of the redistricting data will also be much greater than the demonstration data released to date. By April 30, 2021, the Census Bureau will release a further set of demonstration data that employs a higher privacy-loss budget, tuned for accuracy, and which “better approximates the final privacy-loss budget that will likely be selected for the redistricting data product.” *Abowd Decl.* ¶ 69. Plaintiffs and their experts will have at least four weeks to review the next set of demonstration data, perform their analyses, and submit feedback before DSEP sets the final privacy-loss budget

and production parameters in June. See U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), available [here](#).

Finally, even assuming that the 2018 Operational Plan could be said to violate the APA § 706(2), see Compl. ¶¶ 210–218, the only remedy would be to “set [it] aside” and “remand [it] to the agency for additional investigation.” 5 U.S.C. § 706(2); *Fla. Power & Light Co. v. Lorion*, 470 U.S. 729, 744 (1985). Indeed, under APA § 706(2), “it is not a court’s role to direct the agency how to act. Rather, a court’s role is to review the agency’s decision and, if it cannot be sustained, remand to the agency.” *Neto v. Thompson*, -- F. Supp. 3d --, 2020 WL 7310636, at \*11 (D.N.J. Dec. 10, 2020) (citing *Dep’t of Homeland Sec. v. Regents of the Univ. of California*, 140 S. Ct. 1891, 1907–08 (2020)). And any such remand would add “multiple months” of further delay. Abowd Decl. ¶ 85; see generally *supra*, Background Part D.

#### **4. The Doctrine of Laches Bars Plaintiffs’ Differential Privacy Claims**

Assuming the Court concludes that Plaintiffs are bringing a facial challenge to the 2018 Operational Plan (as opposed to a challenge to the application of differential privacy, which would be premature), such a challenge is barred by the doctrine of laches. The doctrine of laches “protect[s] defendants against unreasonable, prejudicial delay in commencing suit.” *SCA Hygiene Prods. Aktiebolag v. First Quality Baby Prods., LLC*, 137 S. Ct. 954, 960 (2017). The doctrine provides defendants with an equitable defense that warrants consideration “separate from a statute of limitations [defense].” *Grayson v. Allen*, 499 F. Supp. 2d 1228, 1236 (M.D. Ala.), *aff’d*, 491 F.3d 1318 (11th Cir. 2007). The doctrine “will bar a claim when three elements are present: (1) a delay in asserting a right or a claim; (2) that the delay was not excusable; and (3) that there was undue prejudice to the party against whom the claim is asserted.” *Venus Lines Agency, Inc. v. CVG Int’l Am., Inc.*, 234 F.3d 1225, 1230 (11th Cir. 2000); see also *Wood v. Raffensperger*, No. 1:20–CV–04651–

SDG, -- F. Supp. 3d --, 2020 WL 6817513, at \*7 (N.D. Ga. Nov. 20, 2020), *aff'd*, 981 F.3d 1307 (11th Cir. 2020). All three elements are easily satisfied here.

*First*, Plaintiffs have delayed considerably in asserting their claims. Plaintiffs acknowledge that the Bureau announced its decision to use differential privacy for the 2020 Census “in September 2017” and added differential privacy to the 2020 Census Operational Plan “in December 2018.” Mot. 39. Under this theory, Plaintiffs knew or should have known the facts giving rise to their claims by December 2018 at the latest. Rather than timely bringing their claims once Plaintiffs became aware of the Bureau’s plans, however, Plaintiffs waited years to bring their lawsuit, until after the Bureau had already begun processing data and is now on the verge of releasing data in a matter of months. This years-long wait undoubtedly counts as a “delay.” *See, e.g., Wood*, 2020 WL 6817513, at \*7 (laches barred challenge to November 2020 election where plaintiff was aware of basis for claim as early as March 2020); *Stone v. U.S. Postal Serv.*, 383 F. App’x 873, 875 (11th Cir. 2010) (laches barred claim due to plaintiffs’ three-year delay).

*Second*, Plaintiffs’ delay is inexcusable. Plaintiffs take the position that the Census Bureau’s December 2018 operational plan constitutes final agency action that is “ripe for review.” Mot. 39–40. Given that position, there is no excuse for waiting more than two years to challenge that decision. To be sure, the Bureau continues to refine its differential-privacy algorithm, and has not yet set the privacy-loss budget. But in Plaintiffs’ view, that decision is “immaterial” to their claims because “by definition, *any* application of differential privacy will produce erroneous numbers.” *Id.* at 35, 40 (emphasis added). Plaintiffs identify no reason in either their complaint or their motion why they waited until the eleventh hour to file suit. Indeed, Alabama *did* file suit against the Census Bureau in 2018 over the Bureau’s “Residence Rule” – a suit that remains pending in the Northern District of Alabama. *See Compl., Alabama v. Dep’t of Commerce*, No. 18–cv–772 (N.D. Ala. May 21, 2018). But Alabama waited until March 2021 to bring any challenge

to the Bureau's plan to use differential privacy, despite their claim that "any" application of differential privacy would be unlawful.

*Third*, Plaintiffs' delay has unduly prejudiced Defendants. If Plaintiffs had brought their challenge when the Census Bureau announced it would be using differential privacy, the Bureau would have had ample time to implement any operational consequences of an adverse decision before releasing redistricting data to the states. Now, with post-processing operations well underway and the release of data fast approaching, an adverse decision would significantly disrupt the Bureau's completion of the census. As Dr. Abowd explains, it would take "multiple months" to develop, test, and implement an alternative disclosure methodology. Abowd Decl. ¶ 85. Changing course at the last minute also poses significant risks to the accuracy of the data. *See* Thieme Decl. ¶ 74. Moreover, by bringing suit now during what is the busiest time of the decade for the Census Bureau, Plaintiffs have subjected the Bureau to the significant and unnecessary burden of having to defend against a federal lawsuit seeking to upend its entire framework for ensuring privacy while simultaneously working to complete the actual census itself. All of this could have been avoided if Plaintiffs had not delayed in bringing their claims.

**B. Plaintiffs' Challenge to the February 12 Press Release Is Not Likely to Succeed.**

Plaintiffs bring two statutory challenges to the Bureau's February 12 Press Release announcing that it would release redistricting data by September 30, 2021: (i) a claim that the press release "violates the Census Act," Mot. 44-45; Compl. ¶¶ 219-22, and (ii) a claim that the press release violates the APA, Mot. 46-50; Compl. ¶¶ 223-27. Neither challenge is likely to succeed.

**1. Plaintiffs' Claim that the Press Release "Violates the Census Act" Is Not Likely to Succeed**

Plaintiffs are unlikely to succeed on their claim that the February 12 Press Release violates § 141(c) of the Census Act. As an initial matter, Plaintiffs lack a private right of

action to bring this claim. As noted, the only private right of action to enforce § 141(c) flows through § 209(b) of the 1998 Appropriations Act. But that section provides a private right of action only to certain statutorily defined “aggrieved persons” to challenge “the use of any statistical method in violation of the Constitution or any provision of law . . . to determine the population for purposes of the apportionment or redistricting of Members in Congress.” And none of the Plaintiffs can use § 209 to challenge the February 12 Press Release because § 209 allows for challenges only to “statistical methods,” and the press release is obviously not a “statistical method.”<sup>5</sup> Plaintiffs argue that the February 12, 2021 Press Release was “likely” a “byproduct of its . . . decision to implement differential privacy,” which Plaintiffs contend is a “statistical method.” See Pls. Mem., Doc. 2, at 4–5. But Plaintiffs are wrong as a factual matter – as the Thieme declaration explains, the “creation of the [Microdata Detail File] is not the reason that the Census Bureau will be unable to meet the statutory deadline.” See Thieme Decl. ¶ 71. Indeed, the Bureau has allotted approximately three weeks to apply differential privacy, while the disclosure-avoidance procedures used in the 2010 census took nearly four weeks. *Id.* And, more fundamentally, § 209 does not allow for challenges to press releases that are alleged “by-product[s]” of a statistical method – whatever that means. It allows only for challenges to statistical methods themselves.

Plaintiffs thus have no cause of action under the Census Act or § 209 to pursue an alleged violation of the statutory deadline in § 141(c). Nor is there any other basis for Plaintiffs to pursue this claim. While federal courts may “in some circumstances” grant injunctive relief against officials who are alleged to have violated federal law, “[t]he power of federal courts of equity to enjoin unlawful executive action is subject to express and implied statutory limitations.” *Armstrong v. Exceptional Child Ctr., Inc.*, 575 U.S. 320,

---

<sup>5</sup> Additionally, as explained above, Alabama is not an “aggrieved person” under the statute, and so Alabama could not take advantage of § 209(b)’s narrow cause of action to enforce § 141(c) in any event. See *supra* Part II.A.1.

326–27 (2015). By expressly authorizing a cause of action for “aggrieved persons” to bring claims challenging “statistical methods” –but *only* statistical methods—Congress impliedly limited plaintiffs’ ability to challenge other alleged violations of the Census Act. *See id.* at 328 (holding that Medicaid Act foreclosed equitable relief because “sole remedy” Congress provided for in statute was for Secretary to withhold funds).

Nor is review available under the “*ultra vires*” doctrine or any other purported nonstatutory basis for review. Review under the *ultra vires* doctrine “is essentially a Hail Mary pass – and in court as in football, the attempt rarely succeeds.” *Nyunt v. Broad. Bd. of Governors*, 589 F.3d 445, 449 (D.C. Cir. 2009) (Kavanaugh, J.). Among other requirements, a plaintiff must show that the agency’s error is “so extreme that one may view it as jurisdictional or nearly so.” *Id.* (quoting *Griffith v. Fed. Labor Relations Auth.*, 842 F.2d 487, 493 (D.C. Cir. 1988)); *see also Protect Our Parks, Inc. v. Chicago Park Dist.*, 971 F.3d 722, 728 (7th Cir. 2020) (plaintiffs must show that defendants acted “beyond their legal authority”). Plaintiffs have not even attempted to make that showing here. Plaintiffs do not argue that the Census Bureau lacks the statutory authority to report tabulations of population after the deadline has passed, so *ultra vires* review does not even apply. And even if it did, Plaintiffs cannot show that the agency’s error was “so extreme” as to be “jurisdictional or nearly so,” where the Bureau could not meet the statutory deadline due to extraordinary events outside its control.

Finally, even if Alabama had a cause of action under the statute or otherwise, injunctive relief would be inappropriate because, as noted, it is physically impossible for the Bureau to produce redistricting data at this time or any time in the immediate future. A court may not exercise its equitable powers to “require an agency to render performance that is impossible.” *Am. Hosp. Ass’n*, 867 F.3d at 167.

**2. Alabama’s APA Challenge to the February 12 Press Release Is Not Likely to Succeed**

Alabama is likewise unlikely to succeed under the APA because its claim does not challenge any final agency action. Alabama’s claim focuses exclusively on the Bureau’s February 12 Press Release and related blog post. Mot. 44–45 (citing Mot. Exs. 7 & 8). But, as explained above, final agency action occurs when the Secretary reports the final redistricting numbers. *See* Part II.A.3.a.; *Franklin*, 505 U.S. at 790; *City of Detroit*, 4 F.3d at 1377 n.6. So the Press Release is not final agency action reviewable under the APA.

**a. The February 12 Press Release Was Not Final Agency Action**

As explained above, final agency action “must mark the consummation of the agency’s decision-making process—it must not be of a merely tentative or interlocutory nature” and “must be one by which rights or obligations have been determined, or from which legal consequences will flow.” *Bennett*, 520 U.S. at 177–78. A cognizable APA claim must also challenge a “circumscribed, discrete agency action[]”; it cannot advance a “broad programmatic attack” on an agency’s operations. *SUWA*, 542 U.S. at 61–62. Alabama’s challenge to the February 12 Press Release satisfies none of the requirements for final agency action.

*No Consummation of the Decisionmaking Process.* To determine whether an agency action is final, “[t]he core question is whether the agency has completed its decisionmaking process.” *Franklin*, 505 U.S. at 797. The APA does not allow a party to challenge “preliminary, procedural, or intermediate agency action” until the agency completes its action. *See Nat’l Parks Conservation Ass’n*, 324 F.3d at 1236 (quoting 5 U.S.C. § 704).

As explained above, the Supreme Court has held that there is no final agency action until the President delivers the final apportionment count to Congress. *See Franklin*, 505 U.S. at 797. The interim steps taken by the Secretary of Commerce and the Census Bureau prior to the delivery of the final apportionment numbers are tentative and not final agency action. *Id.* Although *Franklin* dealt with apportionment, the same analysis

applies to the redistricting context. *See City of Detroit*, 4 F.3d at 1377 n.6 (relying on *Franklin*'s reasoning to conclude that "the Secretary's reporting of the [redistricting] counts for these purposes is a final agency action"). Since reporting of final redistricting data is reviewable final agency action, the tentative actions and decisions leading up to the delivery of the redistricting data are not reviewable under the APA.

Even setting aside this Supreme Court precedent, a press release explaining that the Census expects to deliver redistricting data by a certain date did not consummate anything; it simply provided a snapshot in time of the expected delivery date that had shifted over the past year due to many factors, including disruptions from COVID, wildfires, hurricanes, court orders, and issues in data processing. *See supra* Background Part E. The February 12 Press Release simply updated Census's estimated timeline, and of course, estimates can still change as data processing continues. *See Whitehorne Decl.* ¶ 17. The Press Release thus does not reflect any definitive decision at all.

*No Legal Consequences.* The February 12 Press Release is also not final agency action because it did not change any legal rights or have any legal consequences. *See Cal. Cmty. Against Toxics v. EPA*, 934 F.3d 627, 638 (D.C. Cir. 2019) (no final agency action where "no direct and appreciable legal consequences" and no party "can rely on it as independently authoritative in any proceeding"). The February 12 Press Release did not change any rights or obligations: the Secretary will deliver redistricting data to the States, including Alabama, when the data becomes available. Like the 2018 Operational Plan, the Press Release was also "purely informational"; "[c]ompelling no one to do anything," the Press Release "had no binding effect whatsoever – not on the agency and not on" the general public. *Indep. Equip. Dealers Ass'n*, 372 F.3d at 427. And, as discussed above, Alabama faces no legal consequences if it does not receive redistricting data by the statutory deadline. *See generally supra* Part I.A.1.b. In fact, Alabama faces no legal consequences at all,



regardless of timing, because its own law fully contemplates how to accomplish apportionment and redistricting in the absence of what it considers to be “full and satisfactory” census data. *See* Ala. Const. § 201; *Ohio*, 2021 WL 1118049, at \*6.

*Improper Programmatic Attack.* Finally, Alabama’s challenge to the February 12 Press Release fails the final-agency-action inquiry because it is a “broad programmatic attack” on the Census Bureau’s operations, not a “circumscribed, discrete agency action[.]” *SUWA*, 542 U.S. at 61–62. While “[c]ourts are well-suited to reviewing specific agency decisions,” they are “woefully ill-suited [ ] to adjudicate generalized grievances asking [them] to improve an agency’s performance or operations.” *City of New York*, 913 F.3d at 431. But that is exactly what Alabama seeks here. Because the Census Bureau’s data-processing operations are all interdependent and interrelated, *see, e.g.*, Thieme Decl. ¶ 5; Whitehorne Decl. ¶¶ 15–16, 21, producing redistricting data on a different timeline would require “a sweeping overhaul to the [processing operations], which exceeds the scope of reviewable ‘agency action.’” *NAACP*, 399 F. Supp. 3d at 422. Indeed, like the Census Bureau’s field operations, its data-processing operations “expressly are tied to one another,” so altering any of these operations “would impact the efficacy of the others, and inevitably would lead to court involvement in ‘hands-on’ management of the Census Bureau’s operations.” *NAACP*, 945 F.3d at 191 (citing *SUWA*, 542 U.S. at 66–67). That is “precisely the result that the ‘discreteness’ requirement of the APA is designed to avoid.” *Id.* (citing *SUWA*, 542 U.S. at 67).

**b. The February 12 Press Release is Not Arbitrary or Capricious**

Nor can Alabama demonstrate that the February 12 Press Release is arbitrary or capricious in violation of the APA. Where (unlike here) there is final agency action, the arbitrary and capricious standard is “exceedingly deferential.” *Sierra Club*, 526 F.3d at 1360. The Court is “not authorized to substitute [its] judgment for the agency’s as long as its conclusions are rational.” *Miccosukee Tribe of Indians of Fla.*, 566 F.3d at 1264. And

this Court should “give an extreme degree of deference to the agency when it is evaluating scientific data within its technical expertise.” *Nat’l Mining Ass’n*, 812 F.3d at 866; see also *Ranchers Cattlemen Action Legal Fund v. Dep’t of Agric.*, 415 F.3d 1078, 1093 (9th Cir. 2005) (“Deference to the informed discretion of the responsible federal agencies is especially appropriate, where, as here, the agency’s decision involves a high level of technical expertise.”).

Here, there is a reasoned explanation for the Secretary’s inability to transmit redistricting data by the statutory deadline: “[I]t is not possible under any scenario for the Census Bureau to produce these data at this time or any time in the immediate future.” Whitehorne Decl. ¶ 14. Nor can the Bureau’s delivery of redistricting data for all States at once be considered arbitrary or capricious. *Contra* Mot. 47. Even if the Census Bureau prioritized Plaintiff’s redistricting data to the detriment of the other 49 States, “it would not be able to deliver the data more than a few weeks earlier than a single national release”; “[t]he resulting data may have uncaught errors from [having] been rushed through review without the benefit of review of all States at once”; and it would “delay the release of data for the other 49 states.” Whitehorne Decl. ¶¶ 29–30.

Finally, even assuming that the February 12 Press Release could be considered “arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law,” the only remedy would be to “set [it] aside” and “remand [it] to the agency for additional investigation.” 5 U.S.C. § 706(2); *Fla. Power & Light Co.*, 470 U.S. at 744. Indeed, under the APA § 706(2), “it is not a court’s role to direct the agency how to act. Rather, a court’s role is to review the agency’s decision and, if it cannot be sustained, remand to the agency.” *Neto*, 2020 WL 7310636, at \*11 (citing *Regents of the Univ. of California*, 140 S. Ct. at 1907–08). And while the Census Bureau would take any such remand seriously, it

would not change the fact that “it is not possible under any scenario for the Census Bureau to produce these data at this time or any time in the immediate future.” Whitehorne Decl. ¶ 14.<sup>6</sup>

**C. Plaintiffs Will Suffer No Harm, Much Less Irreparable Harm.**

“A showing of irreparable injury is the sine qua non of injunctive relief.” *Siegel v. LePore*, 234 F.3d 1163, 1176 (11th Cir. 2000) (en banc) (per curiam). And “the asserted irreparable injury must be neither remote nor speculative, but actual and imminent.” *Id.* “Issuing a preliminary injunction based only on a possibility of irreparable harm is inconsistent with [the Supreme Court’s] characterization of injunctive relief as an extraordinary remedy that may only be awarded upon a clear showing that the plaintiff is entitled to such relief.” *Winter*, 555 U.S. at 22. Here, Plaintiffs cannot establish that they will likely suffer irreparable harm as a result of either the Bureau’s use of differential privacy or its February 12 Press Release.

**1. Plaintiffs Have Not Established Irreparable Harm Due to Differential Privacy**

As a threshold matter, and assuming that the Court concludes that Plaintiffs are bringing a facial challenge to the 2018 Operational Plan (because any challenge to the application of differential privacy is premature), Plaintiffs’ unexplained delay in bringing their differential privacy claim undercuts their claim of irreparable injury. “[T]he very idea of a preliminary injunction is premised on the need for speedy and urgent action to protect a plaintiff’s rights before a case can be resolved on its merits.” *Wreal, LLC v. Amazon.com, Inc.*, 840 F.3d 1244, 1248 (11th Cir. 2016) (emphasis in original). “For this reason” federal courts “have found that a party’s failure to act with speed or urgency in

---

<sup>6</sup> Contrary to Alabama’s protestations, Mot. 47, the Census Bureau *did* consider States’ self-imposed reliance on census-based redistricting data. As the Whitehorne declaration explains, however, “[w]ith the delay in the delivery of the redistricting data, there are now too many states (at least 27) to prioritize, in a fair, logical, and data-driven manner.” Whitehorne Decl. ¶ 26.

moving for a preliminary injunction necessarily undermines a finding of irreparable harm.” *Id.* Thus in *Wreal*, the Eleventh Circuit stated that “[a] delay in seeking a preliminary injunction of even only a few months—though not necessarily fatal—militates against a finding of irreparable harm.” *Id.*

The record here reflects Plaintiffs’ unexplained delay of at least two years. Plaintiffs represent in their motion that the Census Bureau announced its decision to use differential privacy in September 2017, and that the Census Bureau added differential privacy to its “fourth (and latest) version of the Bureau’s 2020 Census Operational Plan,” which was released in December 2018. Mot. 12. They reference demonstration data that the Census Bureau released in October 2019 and in May, September, and November of 2020 that, in their view, “have shown that differential privacy . . . inhibits a State’s right to draw fair lines.” *Id.* at 18. And though the Census Bureau continues to refine its differential-privacy algorithm and its various inputs, Plaintiffs’ position is that “by definition, *any* application of differential privacy will produce erroneous numbers.” *Id.* at 35 (emphasis added).

But Plaintiffs do not explain why they failed to bring a challenge shortly after the Census Bureau added differential privacy to its December 2018 operational plan. Nor do they explain why they didn’t bring such a challenge after the Census Bureau started releasing demonstration data in October 2019. Instead, for reasons they do not explain, Plaintiffs waited until March 2021 to file this suit and move for a preliminary injunction. “[A] party cannot delay . . . and then use an ‘emergency’ created by its own decisions concerning timing to support a motion for a preliminary injunction.” *Mortensen v. Mortg. Elec. Registration Sys., Inc.*, No. CV 09-0787-WS-N, 2010 WL 11425328, at \*8 (S.D. Ala. Dec. 23, 2010). “[B]ecause the instant motion for preliminary injunction was filed not just months, but years, after the factual basis of the Plaintiffs’ claims were known to them, the Plaintiffs have not shown they will suffer imminent, irreparable harm.” *Thompson v. Merrill*, No. 2:16-cv-783-ECM, 2020 WL 3513497, at \*3 (M.D. Ala. June 29, 2020) (Marks, C.J.).

Setting aside Plaintiffs' unexplained delay in bringing their claim, Plaintiffs also cannot demonstrate an irreparable injury because they have not demonstrated any injury at all. *See supra* Part I.A. Plaintiffs contend that they will suffer an irreparable injury because differential privacy will supposedly "make lawful redistricting difficult." Mot. 50. But, as explained above, the redistricting data that the Secretary produces will be perfectly suitable for redistricting. *See* Abowd Decl. ¶¶ 54–56, 65–66, 69. As Dr. Abowd explains, the latest demonstration data product that will be released by April 30 is "extremely accurate." *Id.* ¶ 54. For example, "[t]otal populations for counties have an average error of +/- 5 persons" (an error rate of about 0.04% of the counties' population), whereas "the average county-level estimation uncertainty of the census is +/- 960 persons (averaging 1.6% of the county census counts)." *Id.* "In the April 2021 Demonstration Data Product, Congressional districts as drawn in 2010 have a mean absolute percentage error of 0.06%." *Id.* ¶ 56. And the average state legislative district has an average error of 0.16% or less. *See id.* Such miniscule error cannot possibly interfere with Alabama's ability to "lawful[ly] redistrict[]" or "subject the State to the risk of litigation and liability." Mot. 50. And even if Alabama believed that it did, Alabama's constitution does not require it to use census data in drawing its districts. *See supra* Part I.A.1.b.

Nor have Plaintiffs demonstrated that differential privacy will impose irreparable "financial harm" on Alabama. *See* Mot. 52–55. Again, as explained above, Plaintiffs do not allege that Alabama is likely to suffer a *loss* of federal funds as a result of differential privacy, and make no effort to show that the level of noise that the differential-privacy algorithm will inject will affect any aspect of Alabama's federal funding. *See supra* Part I.A.1.c. To the contrary, Plaintiffs' own expert suggests that to the extent Alabama's funding would be affected by differential privacy at all, it would result in a windfall to the State because, he predicts, rural areas would tend to gain population. *Id.*

Moreover, even if Plaintiffs could establish some potential future injury, they cannot show that they are likely to suffer the kind of "imminent" irreparable harm that

would justify the extraordinary remedy of a preliminary injunction. *Wreal*, 840 F.3d at 1248. As explained above, the Census Bureau is still in the process of finalizing the differential privacy algorithm, and has not, for example, set the privacy-loss budget. *See supra* Background Part D. Until it does so, Plaintiffs cannot demonstrate that the amount of noise that differential privacy adds could possibly be so great as to cause the kinds of irreparable harms that Plaintiffs allege. *See* Mot. 50.

## **2. Plaintiffs Have Not Established Irreparable Harm on Their Delay Claim**

Nor have Plaintiffs demonstrated that they will suffer irreparable harm if the Census Bureau releases redistricting data by September 30, 2021. *See* Mot. 55–56. Again, Plaintiffs have not demonstrated any harm at all, let alone irreparable harm. Plaintiffs’ claim to harm rests entirely on an assertion that Alabama will be unable to comply with its constitution but, as explained above, Alabama’s constitution does not require using decennial census data for redistricting where, as here, the State does not believe that data to be “full and satisfactory.” *See supra* Part I.A.1.b; Ala. Const. § 201. This case is therefore unlike *Maryland v. King*, 567 U.S. 1301 (2012) (Roberts, C.J., in chambers), where a portion of state law was enjoined, precluding the state from enforcing its provisions. *Id.* at 1303 (noting that inability to “employ a duly enacted statute” constitutes irreparable harm). Here, by contrast, Alabama’s constitution expressly contemplates a situation where census data would not be “full and satisfactory” to the State and affords its legislature an opportunity to conduct its own census. *See* Ala. Const. § 201. The realization of a circumstance expressly accounted for in a state’s law is not a frustration of that text or its purpose. *See Conn. Nat’l Bank*, 503 U.S. at 253–54 (courts “must presume that [the] legislature says in a statute what it means and means in a statute what it says there.”).

Alabama may well prefer to use census data for redistricting, but a frustration of an alleged preference, without a factual showing of likely real-world effects, is insufficient to constitute an irreparable injury. *Cf. Judicial Watch, Inc. v. U.S. Dep’t of Homeland*

*Sec.*, 514 F. Supp. 2d 7, 10 (D.D.C. 2007) (“Although plaintiff’s desire to have its case decided in an expedited fashion is understandable, that desire, without more, is insufficient to constitute the irreparable harm[.]”). Were it otherwise, anyone that came to court with a preference for different census operations could obtain an injunction as a matter of course. That is not—and cannot be—the standard. *Siegel*, 234 F.3d at 1179 (“[P]roof of irreparable injury is an indispensable prerequisite to a preliminary injunction.”). And even assuming that Alabama would sustain likely real-world effects, the State has not explained why, unlike other States, *see supra* Background Part E, it cannot find a workable solution other than through this lawsuit.

Likewise, Plaintiffs cannot establish imminent irreparable harm based on the argument that delivering redistricting data by September 30 would leave Alabama’s Boards of Registrars with “only” four months to reassign voters to their correct precincts and districts. Mot. 56. Plaintiffs assert that four months will “likely” not be enough, *id.*, but the declaration that Plaintiffs cite does not support that assertion. *See Helms Decl.*, Doc. 3-3, ¶¶ 5-15. The declaration states merely that in those counties that assign voters manually, the process “can” take “up to [six] months.” *Id.* ¶ 7. This statement appears to be based on one prior reassignment process in 2017 when local officials allegedly struggled to assign voters in six months. *Id.* ¶ 8. From this fact, the declarant infers that requiring officials to complete the reassignment process in four months instead of six “could” lead to increased costs, the “potential[.]” for mistaken reassignments, and the “potential[.]” for confusion. *Id.* ¶ 12. But such “remote [and] speculative” potential harms are insufficient to establish the “actual and imminent” harm necessary to justify a preliminary injunction. *Siegel*, 234 F.3d at 1176. Moreover, Plaintiffs’ declarant acknowledges that Alabama could simply move its 2022 primary election seven weeks to July 12, 2022, *Helms Decl.*, Doc. 3-3, ¶¶ 14-15, which would give Alabama the six months that it says it needs to complete the reassignment process.

Finally, Representative Aderholt cannot establish irreparable harm based on the fact that the Bureau's delay "effectively reduc[es] by at least four months the amount of time [he] can spend campaigning and fundraising." Mot. 56. As explained above, delayed redistricting affects all candidates, and, as the incumbent, Representative Aderholt is perhaps even more likely to benefit from a shorter campaign cycle. *See supra* Part I.A.2. Thus, Representative Aderholt cannot demonstrate any injury at all, let alone an injury that is "actual and imminent."

**D. Defendants and the Public Would Be Harmed by an Injunction.**

*Differential Privacy Is In The Public Interest.* The harm to the government and the public would be severe if the Census Bureau were forced to abandon differential privacy. *See Swain*, 961 F.3d at 1293 (harm to opposing party and the public interest "merge" when relief is sought against the government).

Forcing the Census Bureau to develop a different disclosure-avoidance method would have cascading affects, including significant delay in releasing the redistricting data and decreased quality of the data ultimately released. The Census Bureau is in the final stages of planning how it will deploy differential privacy, which will be the culmination of a process that has been ongoing since at least 2017. Forcing the Bureau to change methods at this late hour would upend the schedule and cause significant delays – indeed, changing methods "would add significant additional time (at least several months) to the schedule for delivering redistricting data." Thieme Decl. ¶ 74. Since the Bureau announced that it would use differential privacy in 2017, States and other data users have provided "extensive actionable feedback" that "has informed ongoing [disclosure-avoidance] system improvements and design changes." U.S. Census Bureau, *2020 Disclosure Avoidance System Updates* (Feb. 23, 2021), available [here](#). Only one State – Alabama – has filed a lawsuit over the use of differential privacy. The other States deserve to get the data they expect without additional, undue delay caused by a preliminary injunction.



There is a strong public interest in protecting the confidentiality of census responses. The Supreme Court has recognized that “an accurate census depends in large part on public cooperation” and “[t]o stimulate that cooperation Congress has provided assurances that information furnished to the Secretary by individuals is to be treated as confidential.” *Baldrige*, 455 U.S. at 354. And a federal statute provides that that Census Bureau staff that publish information protected by 13 U.S.C. § 9 “shall be” subject to fines “or imprisoned not more than 5 years, or both.” 13 U.S.C. § 214.

The Census Bureau chose to use differential privacy because it is the best way to protect confidentiality while still providing quality, accurate redistricting data to the public. Other available disclosure-avoidance methods, including suppression or swapping, do not provide similarly powerful confidentiality protections, and “to achieve the necessary level of privacy protection, both enhanced data swapping and suppression [would have] severely deleterious effects on data quality and availability.” *Abowd Decl.* ¶ 51. And if the Bureau were nonetheless forced to provide detailed data at small geographic levels, it would expose the confidential information of millions of Americans who trusted the Bureau to keep their data secure.

*The Census Bureau Cannot Provide Redistricting Data By March 31, 2021.* It is now April, so it would be impossible for the Bureau to comply with any order requiring it to release redistricting data by March 31, 2021. Even an order requiring the Census Bureau to speed up the release of redistricting data faster than what Census Bureau officials have already announced would be difficult, if not impossible, to implement. *Whitehorne Decl.* ¶¶ 14–17, 21; *see supra* Part I.C. The Census Bureau’s current schedule reflects the realistic amount of time the Bureau has concluded it needs to complete the complex steps required to finish processing the various sources of data it received; verifying the quality of its tabulations; and preparing usable, accurate outputs that comply with statutory requirements for respondent confidentiality protection. *Whitehorne Decl.* ¶¶ 20–21, 28–30; *Thieme Decl.* ¶¶ 60–83 (detailing the steps that still need to be accomplished to deliver

redistricting data). An order requiring the Census Bureau to deliver data faster would yet again disrupt census operations, reduce the time for data quality checks, and make it even *more* difficult for the Census Bureau to complete its work. Whitehorne Decl. ¶¶ 28–30; Thieme Decl. ¶¶ 69, 73–74.

The harm from such a disruption would reverberate to other States and the public at large. If the Census Bureau were required to prioritize Alabama’s data, it may well have to delay delivery of other States’ data until past September 30, 2021. Whitehorne Decl. ¶¶ 30, 31. Such a delay would disrupt those other States’ redistricting plans – presumably leading those States to suffer the same kinds of harms Alabama alleges in this lawsuit. Already, at least one other state has brought a lawsuit like Alabama’s, requesting that its data be prioritized over those of other states. *See Ohio v. Raimondo*, No. 3:21-cv-064, 2021 WL 1118049 (S.D. Ohio Mar. 24, 2021), *appeal filed*, No. 21-3294 (6th Cir. docketed Mar. 25, 2021). Meanwhile, plaintiffs in California continue to assert that any shortening of data-processing operations would be unlawful. *See Nat’l Urban League v. Raimondo*, No. 20-cv-05799, ECF Nos. 465 & 467 (N.D. Cal. Feb. 3, 2021). The more courts intrude on census operations, the more entities will want to seek judicial intervention on their behalf, and the longer it will ultimately take to receive the results.

### III. MANDAMUS RELIEF IS UNAVAILABLE.

In three short paragraphs, Plaintiffs argue that Alabama is entitled to “partial relief through a writ of mandamus requiring the Secretary to meet the statutory deadline of March 31 to deliver the tabulations of populations for redistricting to the States.” Mot. 58–59. “Mandamus is an extraordinary remedy which should be utilized only in the clearest and most compelling of cases.” *Cash v. Barnhart*, 327 F.3d 1252, 1257 (11th Cir. 2003). This is not that case. Plaintiffs’ bid to invoke the Mandamus Act, 28 U.S.C. § 1361, should be rejected.

“Under 28 U.S.C. § 1361, otherwise known as The Mandamus Act, the district court has original jurisdiction over a mandamus action to compel an officer or employee

of the United States or any agency thereof to perform a duty owed to the plaintiff.” *Cash*, 327 F.3d at 1257. “Mandamus relief is appropriate only when: (1) there is no other adequate remedy and (2) the plaintiff has a clear right to the relief requested (in other words, the defendant must have a clear duty to act).” *United States v. Salmona*, 810 F.3d 806, 811 (11th Cir. 2016). “Put another way, a writ of mandamus is intended to provide a remedy for a plaintiff only if he has exhausted all other avenues of relief and only if the defendant owes him a clear nondiscretionary duty.” *Id.* And “[a]lthough the issuance of a writ of mandamus is a legal remedy, it is largely controlled by equitable principles and its issuance is a matter of judicial discretion.” *Cash*, 327 F.3d at 1257–58; *see also, e.g., Lovitky v. Trump*, 949 F.3d 753, 759 (D.C. Cir. 2020) (“Even when the legal requirements for mandamus jurisdiction have been satisfied, however, a court may grant relief only when it finds compelling equitable grounds.”); Mot. 58 (acknowledging that “issuance of the writ” must be “‘appropriate under the circumstances’”) (quoting *Cheney v. United States Dist. Ct.*, 542 U.S. 367, 381 (2004)). Alabama is not entitled to mandamus relief for two independent reasons.

For starters, Alabama has not demonstrated a clear, mandatory duty that would afford it with a clear right to relief because “it is anything but clear that Congress intended the deadline[] at issue to be mandatory rather than directory.” *Friends of Aquifer, Inc. v. Mineta*, 150 F. Supp. 2d 1297, 1300 (N.D. Fla. 2001). Again, mandamus relief presupposes, *inter alia*, that “the defendant owes [the plaintiff] a clear nondiscretionary duty.” *Salmona*, 810 F.3d at 811. And “[f]or there to be a ‘duty owed to the plaintiff’ within the meaning of section 1361, there must be a mandatory or ministerial obligation. If the alleged duty is discretionary or directory, the duty is not ‘owed.’” *Maczko v. Joyce*, 814 F.2d 308, 310 (6th Cir. 1987). To be sure, as Plaintiffs point out, *see* Mot. 44–45, “the word ‘shall’ usually connotes a requirement.” *Maine Cmty. Health Options v. United States*, 140 S. Ct. 1308, 1320 (2020) (emphasis added). But, as the Supreme Court expressly noted, that is not always the case, and it is not the case here.

The *Friends of Aquifer* case is directly on point. That case concerned the Pipeline Safety Act, which provided in part that the Secretary of Transportation “shall prescribe standards” relating to certain hazardous liquid pipeline facilities by various dates certain. 150 F. Supp. 2d at 1298–99 (quoting Pipeline Safety Act, 49 U.S.C. § 60109). The Secretary allegedly did not discharge his statutory duties in that regard, and the plaintiff sought mandamus relief. *See id.* at 1298. Citing several cases, the court explained that “in a variety of contexts, courts have concluded that Congress’s use of the word ‘shall’ in directing the discharge of a specified duty does not require that the statute be construed as mandatory rather than directory.” *Id.* at 1300. The court noted that, like § 141(c) here, the Pipeline Safety Act neither imposed any “penalty or sanction for the Secretary’s failure to prescribe the requisite standards by the specified dates,” nor did it include any provision affording jurisdiction to plaintiffs “to compel the Secretary to prescribe certain standards required under the Act.” *Id.* at 1299–1300. Finding no “clear mandate from Congress that it intended the statutory deadlines at issue to be something other than directory, and absent a showing that Congress intended a clear right in Plaintiff to the relief sought,” the court declined to “exercise its equitable powers to order the Secretary to issue standards that are dependent upon technological complexities and developments that are peculiarly within the agency’s—not th[at] court’s—expertise.” *Id.* at 1301.

The same analysis applies here. Plaintiffs have not demonstrated any “clear mandate from Congress,” *id.*, that it intended the § 141(c) deadline to be mandatory rather than directory. To the contrary, there are no statutory consequences for missing the deadline, and historical practice supports the conclusion that census deadlines are directory in nature. And, like the *Friends of Aquifer* court, this Court should decline to “exercise its equitable powers” to order Defendants to rush the processing of the data Alabama seeks, which work is similarly “dependent upon technological complexities and developments that are peculiarly within” the Census Bureau’s expertise. *See Friends of Aquifer*, 150 F. Supp. 2d at 1301; *see also, e.g., Robertson v. Attorney General of U.S.*, 957 F. Supp. 1035, 1037

(N.D. Ill. 1997) (finding statutory deadline to be directory and declining to issue mandamus relief; “In order to achieve the goals of the statute, the Attorney General and INS may have to engage in lengthy investigations to determine the validity of a given marriage.”).<sup>7</sup>

Moreover, Alabama is not entitled to mandamus relief because, as explained above, the relief it seeks is impossible to provide. “[T]he writ of mandamus will not issue to compel the performance of that which cannot be legally accomplished.” *Am. Hosp. Ass’n*, 867 F.3d at 167. “[P]ossibility is a necessary and antecedent condition for the writ’s issuance.” *Id.* at 169 (collecting sources); see 52 Am. Jur. 2d § 24 (“Mandamus will not issue if the performance of the requested action is impossible”); 55 C.J.S. Mandamus § 19 (“The writ of mandamus will not lie where performance of the duty is impossible.”). Simply put, this Court “may not require” the Census Bureau “to render performance that is impossible.” *Am. Hosp. Ass’n*, 867 F.3d at 167.

This action plainly does not constitute the “the clearest and most compelling of cases” in which to invoke relief under the Mandamus Act. *Cash*, 327 F.3d at 1257. So Plaintiffs’ request for a writ of mandamus must be denied.

## CONCLUSION

For the reasons explained above, Plaintiffs’ motion and petition should be denied.

---

<sup>7</sup> Historical practice demonstrates that Congress considers census deadlines as directory. From the very first census, deadlines were missed for various reasons, but Congress either retroactively revised the statute to accommodate the late submission, or simply ignored that a deadline was missed. See An Act granting further Time for making Return of the Enumeration of the Inhabitants in the District of South Carolina, 1 Stat. 226 (1791). Congress likewise extended census deadlines throughout the 1800s whenever they were missed. See An Act to Extend the Time for Completing the Third Census, 2 Stat. 658 (1811); An Act to Amend the Act Entitled “An Act to Provide for Taking the Fourth Census,” 3 Stat. 643 (1821), An Act to Amend the Act for Taking the Fifth Census, 4 Stat. 439 (1831), An Act to Amend the Act Entitled “An Act to Provide for Taking the Sixth Census,” 5 Stat. 452 (1841), An Act Supplementary to the Act Entitled “An Act Providing for the Taking of the Seventh and Subsequent Censuses,” 9 Stat. 445 (1850).

DATED: April 13, 2021

Respectfully submitted,

BRIAN M. BOYNTON  
Acting Assistant Attorney General

ALEXANDER K. HAAS  
Director, Federal Programs Branch

BRAD P. ROSENBERG  
Assistant Director, Federal Programs Branch

/s/ Elliott M. Davis  
ZACHARY A. AVALLONE  
ELLIOTT M. DAVIS (N.Y. Reg. No. 4596755)  
JOHN ROBINSON  
Trial Attorneys  
Civil Division, Federal Programs Branch  
U.S. Department of Justice  
1100 L St. NW  
Washington, DC 20005  
Phone: (202) 514-5336  
E-mail: [elliott.m.davis@usdoj.gov](mailto:elliott.m.davis@usdoj.gov)

*Counsel for Defendants*

**CERTIFICATE OF SERVICE**

I hereby certify that on April 13, 2021, I filed with the Court and served on opposing counsel through the CM/ECF system the foregoing document.

DATED: April 13, 2021

/s/ Elliott M. Davis  
ELLIOTT M. DAVIS  
(N.Y. Reg. No. 4596755)  
Trial Attorney  
Civil Division, Federal Programs Branch  
U.S. Department of Justice  
1100 L St. NW  
Phone: (202) 514-4336  
Fax: (202) 616-8470  
E-mail: [elliott.m.davis@usdoj.gov](mailto:elliott.m.davis@usdoj.gov)

*Counsel for Defendants*

**IN THE UNITED STATES DISTRICT COURT  
FOR THE MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION**

STATE OF ALABAMA, *et al.*,

Plaintiffs,

v.

UNITED STATES DEPARTMENT OF  
COMMERCE, *et al.*,

Defendants.

Case No. 3:21-CV-211-RAH-ECM-KCN

**DECLARATION OF JOHN M. ABOWD**



I, John M. Abowd, make the following Declaration pursuant to 28 U.S.C. § 1746, and declare that under penalty of perjury the following is true and correct to the best of my knowledge:

**BACKGROUND**

1. I am the Chief Scientist and Associate Director for Research and Methodology at the United States Census Bureau. I have served in this capacity since June 2016. My statements in this declaration are based on my personal knowledge or on information supplied to me in the course of my professional responsibilities.
2. I received my Ph.D. in economics from the University of Chicago with specializations in econometrics and labor economics in 1977 (M.A. 1976). My B.A. in economics is from the University of Notre Dame.
3. I have been a university professor since 1976 when I was appointed assistant professor of economics at Princeton University. I was also assistant and associate professor of econometrics and industrial relations at the University of Chicago Graduate School of Business. In 1987, I was appointed associate professor of industrial and labor relations with indefinite tenure at Cornell University where I am currently the Edmund Ezra Day Professor. I am on unpaid leave from Cornell University to work in my current position at the Census Bureau as part of the Career Senior Executive Service.
4. I am a member and fellow of the American Association for the Advancement of Science, American Statistical Association, Econometric Society, and Society of Labor Economists (president 2014). I am an elected member of the International Statistical Institute. I am also a member of the American Economic Association, International Association for Official Statistics, National Association for Business Economists, American Association for Public Opinion Research, Association for Computing Machinery, and American Association of Wine Economists. I regularly attend and present papers at the meetings of these organizations.

5. I have served on the American Economic Association Committee on Economic Statistics. I have also served on the National Academy of Sciences Committee on National Statistics, the Conference on Research in Income and Wealth Executive Committee, and the Bureau of Labor Statistics Technical Advisory Board for the National Longitudinal Surveys (chair: 1999-2001).
6. I have worked with the Census Bureau since 1998, when the Census Bureau and Cornell University entered into the first of a sequence of Intergovernmental Personnel Act agreements and other contracts. Under those agreements, I served continuously as Distinguished Senior Research Fellow at the Census Bureau until I assumed my current position as Chief Scientist in 2016, under a new Intergovernmental Personnel Act contract. Since March 29, 2020, I have been in the Associate Director position at the Census Bureau as a Career Senior Executive Service employee.
7. From 2011 until I assumed my position as Chief Scientist at the Census Bureau in 2016, I was the lead Principal Investigator of the Cornell University node of the NSF-Census Research Network, one of eight such nodes that worked collaboratively with the Census Bureau and other federal statistical agencies to identify important theoretical and applied research projects of direct programmatic importance to the agencies. The Cornell node produced the fundamental science explaining the distinct roles of statistical policymakers and computer scientists in the design and implementation of differential privacy systems at statistical agencies.
8. I have published more than 100 scholarly books, monographs, and articles in the disciplines of economics, econometrics, statistics, computer science, and information science. I have been the principal investigator or co-principal investigator on 35 sponsored research projects. I was a founding editor of the Journal of Privacy and Confidentiality – an interdisciplinary journal, and I continue to serve as an editor and on the governance board. My full professional resume is attached to this report as Appendix A.

9. I have worked on and managed Census Bureau projects that were precursors to the Census Bureau's current program to implement differential privacy for the 2020 Census of Population and Housing. I was one of three senior researchers who founded the Longitudinal Employer-Household Dynamics (LEHD) program at the Census Bureau, which is generally acknowledged as the Census Bureau's first 21<sup>st</sup> Century data product: built to the specifications of local labor market specialists without additional survey burden, and published beginning in 2001 using state-of-the-art confidentiality protection via noise infusion. This program produces detailed public-use statistical data on the characteristics of workers and employers in local labor markets using large-scale linked administrative, census, and survey data from many different sources. In 2008, my work with LEHD led to the first production implementation worldwide of differential privacy as part of a product of the LEHD program called OnTheMap. The LEHD program also implemented other prototype systems to protect confidential information, including allowing the public to access synthetic micro-data confirmed via direct analysis of the confidential data on validation servers. A differentially private version of this system is under development at the Census Bureau but not for use with the 2020 Census.

#### **IMPORTANCE OF CONFIDENTIALITY**

10. Though participation in the census is mandatory under 13 U.S. Code § 221, in practice, the Census Bureau must rely on the voluntary participation of each household in order to conduct a complete enumeration.
11. One of the most significant barriers to conducting a complete and accurate enumeration are individuals' concerns about the confidentiality of census data. The Census Bureau's pre-2020 Census research showed that 28% of respondents were "extremely concerned" or "very concerned" and a further 25% were "somewhat concerned"

about the confidentiality of their census responses.<sup>1</sup> These concerns are even more pronounced in minority populations and represent a major operational challenge to enumerating traditionally hard-to-count populations.<sup>2</sup>

12. To secure voluntary participation, Congress first established confidentiality protections for individual census responses in the Census Act of 1879. These confidentiality protections were later expanded and codified in 13 U.S. Code §§ 8(b) & 9, which prohibits the Census Bureau from releasing “any publication whereby the data furnished by any particular establishment or individual under this title can be identified[.]” and allows the Secretary to provide aggregate statistics so long as those data “do not disclose the information reported by, or on behalf of, any particular respondent[.]” Title III of the Foundations for Evidence Based Policymaking Act of 2018 also requires statistical agencies to “protect the trust of information providers by ensuring the confidentiality and exclusive statistical use of their responses.”<sup>3</sup>
13. The broader scientific community generally concurs about the importance of rigorous protection of confidentiality by statistical agencies. For example, the National Academy of Sciences’ definitive guidebook for federal statistical agencies states “Because virtually every person, household, business, state or local government, and organization is the subject of some federal statistics, public trust is essential for the continued effectiveness of federal statistical agencies. Individuals and entities providing data di-

---

<sup>1</sup> U.S. Census Bureau (2019) “2020 Census Barriers, Attitudes, and Motivators Study Survey Report” <https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-cbams-study-survey.pdf>, p.38-39.

<sup>2</sup> Ibid, p.39-42.

<sup>3</sup> Title III of the Foundations for Evidence Based Policymaking Act of 2018, § 3563.

rectly or indirectly to federal statistical agencies must trust that the agencies will appropriately handle and protect their information.”<sup>4</sup> The report also notes that respondents expect statistical agencies not to “release or publish their information in identifiable form.”<sup>5</sup> The National Academies also broadly exhort statistical agencies to “continually seek to improve and innovate their processes, methods, and statistical products to better measure an ever-changing world.”<sup>6</sup>

14. The Census Bureau enjoys higher self-response rates than private survey companies in large part because the public generally trusts the Census Bureau to keep its data safe. The Census Bureau makes extensive outreach efforts to assure respondents and other data providers about the Bureau’s commitment to protection of confidential data. The criminal fines and imprisonment penalties that Census Bureau employees would face by unlawfully disclosing respondent information are frequently cited by the Census Bureau in these outreach efforts.<sup>7</sup>
15. This trust in the Census Bureau is particularly important for the decennial census, given the “civic ceremony” aspect of the census, akin to the civic ceremony aspect of elections and voting. The decennial census is an exercise where the nation comes together every ten years, under a strict promise of confidentiality, to provide information to help govern our nation. Were the Census Bureau to expose confidential information, there is no doubt that self-response rates would drop, increasing survey

---

<sup>4</sup> National Academies of Sciences, Engineering, and Medicine 2021. Principles and Practices for a Federal Statistical Agency: Seventh Edition. Washington, DC: The National Academies Press. <https://doi.org/10.17226/25885>, p. 37-38.

<sup>5</sup> Ibid., p.38.

<sup>6</sup> Ibid., p.4.

<sup>7</sup> <https://www.census.gov/content/dam/Census/library/factsheets/2019/comm/2020-confidentiality-factsheet.pdf>.

cost across programs by increasing in-person follow up, and decreasing the quality of the census overall.

#### **PRIVACY PROTECTION AT THE CENSUS BUREAU**

16. Protecting privacy is at the core of the Census Bureau’s mission. Our privacy promise to respondents is key to promoting response to our censuses and surveys. The Census Bureau – at the crux of its dual mandate to publish only statistical summaries and to protect the confidentiality of respondent data – is balancing the preferences of data users and data providers. An optimal choice must account for the preferences of data users and protect the data the American people entrust the Census Bureau with keeping safe.<sup>8</sup>
17. Data collected from the decennial census support a wide array of critical government and societal functions at the federal, state, tribal, and local levels. In addition to apportioning seats in the U.S. House of Representatives and supporting the redistricting of those seats, census data also support the allocation of over \$675 billion in federal

---

<sup>8</sup> “Official Statistics at the Crossroads: Data Quality and Access in an Era of Heightened Privacy Risk,” *The Survey Statistician*, 2021, Vol. 83, 23-26 (available at [Survey Statistician 2021 January N83 03.pdf \(isi-iass.org\)](https://www.isi-iass.org/Survey-Statistician-2021-January-N83-03.pdf)). The paper is based on talks that I gave in 2019 to the Committee on National Statistics and the Joint Statistical Meetings. It summarizes the research in Abowd, J.M. and I. Schmutte “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices,” *American Economic Review*, Vol. 109, No. 1 (January 2019):171-202, DOI:[10.1257/aer.20170627](https://doi.org/10.1257/aer.20170627).

funding each year based on population counts, geography, and demographic characteristics.<sup>9</sup> Census data also support important public and private sector decision-making at the federal, state, tribal, and local levels, and serve as benchmark statistics for other important surveys and data collections throughout the decade.<sup>10</sup>

18. The Census Bureau publishes an enormous number of statistics calculated from its collected data. Following the 2010 Census, for example, the Census Bureau published over 150 billion independent statistics about the characteristics of the 308,745,538 persons in the resident population that were enumerated in the census. To serve their intended governmental and societal uses, the majority of these statistics needed to be published at very fine levels of detail and with geographic precision often down to the individual census tract or block.
19. While it would be quite difficult from any single one of those published statistics to ascertain the identity of any individual census respondent or the contents of that respondent's census response, the volume and detail of information published by the Census Bureau, taken together, pose a serious challenge for protecting the privacy and confidentiality of census data. Combining information from multiple published statistics or tables can make it easy to pick out those individuals in a particular geographic area whose characteristics differ from those of the rest of their neighbors. These individuals, who have unique combinations of the demographic characteristics

---

<sup>9</sup> Hotchkiss, M., & Phelan, J. (2017). Uses of Census Bureau data in federal funds distribution: A new design for the 21st century. United States Census Bureau. <https://www2.census.gov/programs-surveys/decennial/2020/program-management/working-papers/Uses-of-Census-Bureau-Data-in-Federal-Funds-Distribution.pdf>.

<sup>10</sup> Sullivan, T. A. (2020). Coming to Our Census: How Social Statistics Underpin Our Democracy (and Republic). *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.c871f9e0>.

reported in statistical summaries, are known as “population uniques” and their records have traditionally been the target of the mechanisms that the Census Bureau uses to protect confidentiality in its data publications.

20. Traditional statistical disclosure limitation methods,<sup>11</sup> like those used in 2010 census, cannot defend against modern challenges posed by enormous cloud computing capacity and sophisticated software libraries. That does not mean traditional statistical disclosure limitation methods usually fail – they usually do not fail. But as computer scientists bring their expertise from the field of cryptography to the field of safe data publication, they have exposed significant vulnerabilities in traditional privacy methods. The Census Bureau’s own internal analysis, for example, confirmed that a modern database reconstruction-abetted re-identification attack can reliably match a large number of 2010 census responses to the names of those respondents – a vulnerability that exposed information of *at least* 52 million Americans and potentially up to 179 million Americans.<sup>12</sup> To defend against this known vulnerability, the Census Bureau explored different confidentiality methods that explicitly defend against database reconstruction attacks and concluded that the best tool to protect against this modern attack while also preserving the accuracy and usability of data products comes from the body of scientific work called “differential privacy.”

#### **THE HISTORY OF INNOVATION IN THE DECENNIAL CENSUS**

21. The decennial census, known officially as the *Decennial Census of Population and Housing*, is the flagship statistical product of the U.S. Census Bureau. Though the Census

---

<sup>11</sup> The technical field that addresses confidentiality is known as “statistical disclosure limitation.” At the Census Bureau, it is known as “disclosure avoidance.” It is also called “statistical disclosure control” by some statisticians and “privacy-preserving data analysis” by some computer scientists.

<sup>12</sup> See Appendix B for a summary of the Census Bureau’s simulated reconstruction and re-identification attacks.



Bureau conducts hundreds of surveys every year, the once-every-decade enumeration of the population of the United States, mandated by Article I, Section 2 of the U.S. Constitution, is the single largest and most complex data collection regularly conducted by the United States government. Since the very first U.S. census in 1790, the collection, processing, and dissemination of census data have posed unique challenges and have required the Census Bureau to improve its operations every decade.

22. The challenges faced by the Census Bureau have led to remarkable innovations. Herman Hollerith's electric tabulation machine, developed for the 1890 Census, revolutionized the field of data processing and led Hollerith to form the company that eventually became IBM.<sup>13</sup> To conduct the 1950 Census, the Census Bureau commissioned the development of the first successful civilian digital computer, UNIVAC I.<sup>14</sup> With each passing decade, the Census Bureau develops, tests, and deploys innovations to its statistical methods, field data collection methods, and data processing operations.

23. That spirit of innovation includes the Census Bureau's more recent implementation of cutting-edge privacy protections. Prior to the 1990 Census, the primary mechanism that the Census Bureau employed to protect the confidentiality of individual census responses was to withhold publication of (or "suppress") any table that did not meet certain household, population, or demographic characteristic thresholds. The 1970 Census, for example, suppressed tables reflecting fewer than five households, and would only publish tables of demographic characteristics cross-tabulated by race if

---

<sup>13</sup> [https://www.census.gov/history/www/census\\_then\\_now/notable\\_alumni/herman\\_hollerith.html](https://www.census.gov/history/www/census_then_now/notable_alumni/herman_hollerith.html).

<sup>14</sup> [https://www.census.gov/history/www/innovations/technology/univac\\_i.html](https://www.census.gov/history/www/innovations/technology/univac_i.html).

there were at least five individuals in each reported race category.<sup>15</sup> These suppression routines helped to protect privacy by reducing the detail of data published about individuals who were relatively unique within their communities. By the 1990 Census, however, the Census Bureau transitioned away from suppression methodologies for two reasons: first, data users were dissatisfied with missing details caused by suppression and second, the Bureau realized that the suppression routines it had been using were insufficient to fully protect against re-identification.<sup>16</sup>

24. For the 1990 Census, the Bureau began using a technique known as noise infusion to safeguard respondent confidentiality. Noise infusion helps to protect the confidentiality of published data by introducing controlled amounts of error or “noise” into the data. The goal of noise infusion is to preserve the overall statistical validity of the resulting data while introducing enough uncertainty that attackers would not have any reasonable degree of certainty that they had isolated data for any particular respondent. The noise infusion used in 1990 was a very simple form of data swapping between paired households in a geographic area with similar attributes, and for small

---

<sup>15</sup> Zeisset, P. (1978), “Suppression vs. Random Rounding: Disclosure Avoidance Alternatives for the 1980 Census,” <https://www.census.gov/content/dam/Census/library/working-papers/1978/adrm/Suppression%20vs.%20Random%20Rounding%20Disclosure-Avoidance%20Alternatives%20for%20the%201980%20Census.pdf>.

<sup>16</sup> McKenna, L. (2018), “Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing,” <https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf>, p.6.

block groups the Census Bureau replaced the collected characteristics of households with imputed characteristics.<sup>17</sup>

25. For the 2000 and 2010 censuses, the Census Bureau began to infuse noise using a more advanced “data swapping” method. The Census Bureau first identified households most vulnerable to re-identification—especially households on smaller-population blocks whose residents had differing demographic characteristics from the remainder of their block. While every non-imputed<sup>18</sup> household record in the Census Edited File (CEF) had a chance of being selected for data swapping, records for more vulnerable households (typically those on low-population blocks) were selected with greater probability. Then, the records for all members of those selected households were exchanged with the records of households in nearby geographic areas that matched on key characteristics. For the 2000 and 2010 censuses, those key matching characteristics were (1) the whole number of persons in the household, and (2) the whole number of persons aged 18 or older in the household. These swapping criteria resulted in the total population and total voting age population for each block being held “invariant”—that is, while noise was added to all remaining characteristics, no noise was added to the block-level total population or block-level voting age population

---

<sup>17</sup> Ibid., p. 6-7. An “imputed characteristic” is the prediction of a statistical model used in place of a missing characteristic, when used in standard editing procedures, or in place of a collected characteristic, when used for confidentiality protection.

<sup>18</sup> When a respondent household provides only a count of the number of persons living at that address or when the housing unit population count is itself imputed, the Census Bureau imputes all characteristics: sex, age, race, ethnicity, and relationship to others in the household. Such persons are called “whole-person census imputations” in technical documentation. When a household consists entirely of whole-person census imputation records, it is called an “imputed” household. A “non-imputed” household contains at least one person whose characteristics were collected on the census form for the household.

counts.<sup>19</sup> *The selection and application of these particular invariants is not an innate feature of data swapping; invariants are implementation parameters that can be applied to (or removed from) any counted characteristic under any noise infusion methodology.*

#### **THE PRIVACY PROTECTIONS USED FOR THE 2010 CENSUS ARE NO LONGER SUFFICIENT**

26. While the Census Bureau's confidentiality methodologies for the 2000 and 2010 censuses were considered sufficient at the time, advances in technology in the years since have reduced the confidentiality protection provided by data swapping.
27. Disclosure avoidance has been a recognized branch of statistics since the 1970s, but it has only been since the late 1990s that it has evolved into a distinct scientific field of study in both statistics and computer science. Prof. Latanya Sweeney's 1997 revelation that she had re-identified then Massachusetts Governor William Weld's medical records in a purportedly "deidentified" public database<sup>20</sup> prompted the Census Bureau and many other statistical agencies to re-examine the efficacy of their disclosure avoidance techniques.
28. *Re-identification attacks.* Prior to 2016, disclosure risk assessments usually focused on assessing the vulnerability of microdata releases (data products that contain individual records for all or some of the data subjects deidentified by removing names and addresses), rather than the rules used for aggregated data releases (data compiled and aggregated into tables). Simulated "re-identification attacks" analyze the risk that an external attacker could use individuals' characteristics that are included on a published microdata file (e.g., location, age, and sex) and link those records to a third-

---

<sup>19</sup> Ibid. p. 8-10.

<sup>20</sup> Sweeney, L. (2002). "k-anonymity: a model for protecting privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5); 557-570, also recounted in Ohm, P. (2009) "Broken promises of privacy: Responding to the surprising failure of anonymization." *UCLA L. Rev.* 57: 1701.

party data source (e.g., commercial data or voter registration lists) that contains those characteristics along with the individuals' names and addresses. The resulting rates of "putative" (suspected) and confirmed linkages show the overall degree of vulnerability of the data. If those linkage rates are deemed too large, then additional disclosure avoidance is necessary to mitigate the disclosure risk.

29. The general problem with relying exclusively on re-identification studies to assess disclosure risk is that they can only provide a "best-case" approximation of the underlying disclosure risk of the data. If a real attacker has access to more sophisticated tools (e.g., optimization algorithms or computing power) or to higher quality external data (e.g., with better age and address information) than the tools or data used in the simulated attack, then the real disclosure risk will be substantially higher than what is estimated via the study. This limitation is particularly vexing for statistical agencies that must rely on a "release and forget" approach to data publication, where disclosure avoidance safeguards must be selected without foreknowledge of the better tools and external data that attackers may have at their disposal after the data are published.
30. Re-identification studies also underestimate the risk from releasing aggregated data. The Census Bureau has long relied on re-identification studies to assess the disclosure risk of its microdata releases, but the majority of Census Bureau data products are aggregated data releases. Over the past decade, aggregated data releases have become increasingly vulnerable to sophisticated "reconstruction attacks" that have emerged as computing power has improved and gotten substantially cheaper.

31. *Reconstruction attacks.* The theory behind a “reconstruction attack” is that the release of *any* statistic calculated from a confidential data source will reveal a potentially trivial, but non-zero, amount of confidential information.<sup>21</sup> As a consequence, if an attacker has access to enough aggregated data with sufficient detail and precision, then the attacker may be able to leverage information from each statistic in the aggregated data to reconstruct the individual-level records that were used to generate the published tables. This process is known as a “reconstruction attack,” and it adds a new degree of disclosure vulnerability against which statistical agencies must defend. While the statistical and computer science communities have been aware of this vulnerability since 2003, only over the last few years have computing power and the sophisticated numerical optimization software necessary to perform these types of reconstructions advanced enough to permit reconstruction attacks at any significant scale.
32. The risk of reconstruction and re-identification attacks is real and substantiated. The Census Bureau has been approached by Prof. Sweeney and others who claim that they have identified specific vulnerabilities in our standard disclosure avoidance methodologies.<sup>22</sup> The vulnerabilities in the disclosure avoidance protections for the Census Bureau’s Survey of Income and Program Participation (SIPP) identified by Prof. Sweeney led the Census Bureau to immediately implement permanent changes to the

---

<sup>21</sup> Dinur, I. and Nissim, K. (2003) “Revealing Information while Preserving Privacy” PODS, June 9-12, San Diego, CA. <https://doi.org/10.1145/773153.773173>.

<sup>22</sup> McKenna, L. (2019b). “U.S. Census Bureau Reidentification Studies,” available at <https://www.census.gov/library/working-papers/2019/adrm/2019-04-ReidentificationStudies.html>.

disclosure avoidance rules used for SIPP data, including increased noise infusion and delayed reporting of survey participants' major life events.<sup>23</sup>

33. Statistical releases do not all need to be of the same type, or contain the same data fields, to enable re-identification by reconstruction. For example, a 2015 interagency report published by the National Institute of Standards and Technology (NIST) written by my colleague Simson Garfinkel provided examples of using disparate data sets to reconstruct hidden underlying data.<sup>24</sup> Some of these examples are quoted here:

34. "*The Netflix Prize*: Narayanan and Shmatikov showed in 2008 that in many cases the set of movies that a person had watched could be used as an identifier.<sup>25</sup> Netflix had released a dataset of movies that some of its customers had watched and ranked as part of its "Netflix Prize" competition. Although there was [sic] no direct identifiers in the dataset, the researchers showed that a set of movies watched (especially less popular films, such as cult classics and foreign films) could frequently be used to match a user profile from the Netflix dataset to a single user profile in the Internet Movie Data Base (IMDB), which had not been de-identified and included user names, many of which were real names. The threat scenario is that by rating a few movies on IMDB, a person might inadvertently reveal *all* of the movies that they had watched, since the person's IMDB profile could be linked with the Netflix Prize data."<sup>26</sup> (emphasis in original)

---

<sup>23</sup> McKenna, L. (2019b). p. 2-3.

<sup>24</sup> Garfinkel, S. (2015) "De-Identification of Personal Information," National Institute of Standards and Technology, available at <http://dx.doi.org/10.6028/NIST.IR.8053> at 26-27.

<sup>25</sup> Narayanan, A. and Shmatikov V. "Robust De-anonymization of Large Sparse Datasets," *IEEE Symposium on Security and Privacy* (2008): 111-125.

<sup>26</sup> Garfinkel, S. (2015), p. 26-27.

35. “*Credit Card Transactions*: Working with a collection of de-identified credit card transactions from a sample of 1.1 million people from an unnamed country, Montjoye *et al.* showed that four distinct points in space and time were sufficient to specify uniquely 90% of the individuals in their sample.<sup>27</sup> Lowering the geographical resolution and binning transaction values (*e.g.*, reporting a purchase of \$14.86 as between \$10.00 and \$19.99) increased the number of points required.”<sup>28</sup>
36. “*Mobility Traces*: Montjoye *et al.* showed that people and vehicles could be identified by their “mobility traces” (a record of locations and times that the person or vehicle visited). In their study, trace data from a sample of 1.5 million individuals was processed, with time values being generalized to the hour and spatial data generalized to the resolution provided by a cell phone system (typically 10-20 city blocks).<sup>29</sup> The researchers found that four randomly chosen observations of an individual putting them at a specific place and time was sufficient to uniquely identify 95% of the data subjects.<sup>30</sup> Space/time points for individuals can be collected from a variety of sources, including purchases with a credit card, a photograph, or Internet usage. A similar study performed by Ma *et al.* found that 30%-50% of individuals could be identified with 10 pieces of side information.<sup>31</sup> The threat scenario is that a person who

---

<sup>27</sup> Montjoye, Y-A. et al. “Unique in the shopping mall: On the reidentifiability of credit card metadata,” *Science*, 30 (January 2015) Vol 347, Issue 6221.

<sup>28</sup> Garfinkel, S. (2015), p. 27.

<sup>29</sup> De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1).

<sup>30</sup> *Ibid.*, p. 1-5.

<sup>31</sup> C. Y. T. Ma, D. K. Y. Yau, N. K. Yip and N. S. V. Rao (2013) "Privacy Vulnerability of Published Anonymous Mobility Traces," in *IEEE/ACM Transactions on Networking*, vol. 21, no. 3, pp. 720-733, June 2013, doi: 10.1109/TNET.2012.2208983.



revealed five place/time pairs (perhaps by sending email from work and home at four times over the course of a month) would make it possible for an attacker to identify his or her entire mobility trace in a publicly released dataset. As above, the attacker would need to know that the target was in the data.”<sup>32</sup>

37. The same general principles apply to census data. The difference between census data and the examples above is that census data can be combined in vastly more ways with other information because all the tables published from census data share basic standardized identifiers including location, age, sex, race, ethnicity, and marital status. Even if each of these identifiers is not included in every table, their use and combinations across many different tables creates the disclosure risk. The Census Bureau understood this emerging risk even before the 2010 Census. As field collection for the 2010 Census was first beginning, the Census Bureau had already flagged the heightened disclosure risk of releasing detailed block level population data, even with the 2010 Census swapping mechanism in place.<sup>33</sup> After tracking this growing risk of reconstruction and re-identification attacks for several years, the Census Bureau decided in 2015 to establish a new team to comprehensively evaluate the Census Bureau’s disclosure avoidance methods to determine if they were sufficient to protect against these disclosure risks.<sup>34</sup>

---

<sup>32</sup> Garfinkel, S. (2015), p. 27-28.

<sup>33</sup> During a January 2010 meeting of the Census Bureau’s Data Stewardship Executive Policy (DSEP) Committee, the chair of the Disclosure Review Board voiced her concerns about the 2010 Census swapping mechanism’s ability to adequately protect future censuses, noting specifically the challenge posed by “continuing to release data at the block level, as block populations continue to decrease (e.g., 40% of blocks in North Dakota have only 1 household in them)” Based on this warning, DSEP decided that “the problem of block population size and disclosure avoidance is real, and that it deserves attention in the context of 2020 planning.” DSEP Meeting Record, January 14, 2010. See Appendix C.

<sup>34</sup> DSEP Meeting Record, February 5, 2015. See Appendix D.

## **2010 CENSUS SIMULATED RECONSTRUCTION-ABETTED RE-IDENTIFICATION ATTACK**

38. The results from the Census Bureau's 2016-2019 research program on simulated reconstruction-abetted re-identification attack were conclusive, indisputable, and alarming. Appendix B, attached to this declaration, provides an overview of that simulation and the results. The bottom line is that our simulated attack showed that a conservative attack scenario using just 6 billion of the over 150 billion statistics released in 2010 would allow an attacker to accurately re-identify *at least* 52 million 2010 Census respondents (17% of the population) and the attacker would have a high degree of confidence in their results with minimal additional verification or field work. In a more pessimistic scenario, an attacker with access to higher quality commercial name and address data than those used in our simulated attack could accurately re-identify around 179 million Americans or around 58% of the population.
39. Emerging attack scenarios and our own internal simulated attacks show that were the Census Bureau to use the disclosure avoidance mechanism implemented for the 2010 Census again for the 2020 Census, the results would be vulnerable to reconstruction and re-identification attacks because of the parameters of the swapping mechanism's 2010 implementation: an overall insufficient level of noise, the invariants preserved without noise, and the geographic and demographic detail of the published summary data. The Census Bureau can no longer rely on the swapping implementation used in 2010 if it is to meet its obligations to protect respondent confidentiality under 13 U.S. Code §§ 8(b) & 9. Protecting against new technology-enabled re-identification attacks, while maintaining the high quality of the decennial census data products, requires the implementation of a disclosure avoidance mechanism that is better able to protect against these new, sophisticated vectors of attack.

#### DISCLOSURE AVOIDANCE OPTIONS CONSIDERED FOR THE 2020 CENSUS

40. Faced with this compelling mathematical and empirical evidence of the inherent vulnerability of the 2010 Census swapping mechanism to protect against reconstruction-abetted re-identification attacks, the Census Bureau began exploring the available data protection strategies that it could employ for the 2020 Census. The three methods the Census considered were *Enhanced Data Swapping*, *Suppression*, and *Differential Privacy*.
41. The Census Bureau decided that differential privacy was the best tool after analyzing the various options through the lens of economics. Efficiently protecting privacy can be viewed as an economic problem because it involves the allocation of a scarce resource—confidential information—between two competing uses: public data products and privacy protection. If we produce more accuracy, we will have less privacy, and vice versa. And just like in the classic economic example of the trade-off between producing guns and butter, the tradeoff between privacy and accuracy can be analyzed with a production possibility curve. Our empirical analysis showed that differential privacy offered the most efficient trade-off between privacy and accuracy—our calculations showed that the efficiency of differential privacy dominated traditional methods.<sup>35</sup> In other words, regardless of the level of desired confidentiality, differential privacy will always produce more accurate data than the alternative traditional methods considered by the Census Bureau.
42. *Enhanced Data Swapping*. Enhancing the data swapping mechanism used for the 2010 Census in a manner sufficient to protect against emerging threats like reconstruction

---

<sup>35</sup> See Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1), 171-202.

attacks would have a significant, detrimental impact on data quality. With an estimated 57% of the population<sup>36</sup> known to be unique at the block level, a swapping mechanism that targets vulnerable households for swapping would require significantly higher rates of swapping than were used in 2010 to protect against a reconstruction attack. Implementing swapping in 2020 would also require abandoning the total population and voting-age population invariants that were used in 2010. There are two technical reasons for this. First, at swap rates sufficient to counter the reconstruction of microdata accurate enough to enable large-scale reidentification, it is impossible to find enough paired households with the same number of persons and adults without searching well outside the neighborhood of the original household. Finding swap pairs was a challenge for some states even at the 2010 swap rate. Second, holding the total and adult populations invariant gives the attacker a huge reconstruction advantage—exact record counts in each block for persons and adults. This advantage vastly improves the accuracy of the reconstructed data. Even a small amount of uncertainty about the block location of an individual greatly expands the variability in the reconstructed microdata effectively reducing the chances of a correct linkage in a re-identification attack. If a block is known to contain exactly seven persons in the confidential data, then every feasible reconstructed version of those data will have exactly seven records in that block, meaning that the block identifier will be correct on every record of every feasible reconstructed database. But if the block population is reported with some random fluctuation around seven, then only by chance will the

---

<sup>36</sup> Fifty-seven percent of the 308,745,538 person records in the confidential 2010 Census Edited File, the definitive source for all 2010 Census tabulations, were unique on their block location, sex, age (in years), race (any combination of the 6 OMB-approved race categories, 63 possibilities in all) and Hispanic/Latino ethnicity. This previously confidential statistic was approved for publication with DRB clearance number CBDRB-FY21-DSEP-003.

block identifier be correct in the reconstructed data. Compound this effect over 8,000,000 blocks and the number of feasible reconstructions explodes exponentially. This is what provides the protection against re-identification from the reconstructed data.<sup>37</sup> Internal experiments also confirmed that increasing the swap rate from the level used in 2010 and removing the invariants on block-level population counts (to permit the increased level of swapping and protect against reconstruction attacks) would render the resulting data unusable for most data users.

43. *Suppression*. While the Census Bureau could use suppression to protect from a reconstruction attack, the resulting data would be only available at a very high level of generality. Today's data users, including redistricters, rely on detailed block and tract-level data, which would not be available for many areas if the Census were to return to suppression to protect against modern attacks.
44. *Differential Privacy*. Differential privacy, first developed in 2006, is a framework for quantifying the precise disclosure risk associated with each incremental release from a confidential data source.<sup>38</sup> In turn, this allows an agency like the Census Bureau to quantify the precise amount of statistical noise required to protect privacy. This precision allows the Census to calibrate and allocate precise amounts of statistical noise in a way that protects privacy while maintaining the overall statistical validity of the data.

---

<sup>37</sup> Garfinkel, S., Abowd, J. M., & Martindale, C. (2018). Understanding Database Reconstruction Attacks on Public Data: These attacks on statistical databases are no longer a theoretical danger. *Queue*, 16(5), 28-53.

<sup>38</sup> Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer, Berlin, Heidelberg.

45. The Census Bureau first began using differential privacy to protect its statistical data products in 2008, with the launch of its OnTheMap tool for employee commuting statistics and its heavily used extension OnTheMap for Emergency Management. In the years since, the Census Bureau has also successfully used differential privacy in a number of other innovative statistical products, such as the Post-Secondary Employment Outcomes and Veteran Employment Outcomes products. Differential privacy is also being used by many of the major technology firms, including Apple<sup>39</sup>, Google,<sup>40</sup> Microsoft,<sup>41</sup> and Uber.<sup>42</sup> Other statistical agencies, such as the Statistics of Income Division of the Internal Revenue Service, have also begun implementing differential privacy.<sup>43</sup> Internationally, the Australian Bureau of Statistics,<sup>44</sup> the Office of National

---

<sup>39</sup>Differential Privacy Team. (2017). "Learning with Privacy at Scale." *Apple Machine Learning Journal*, 1(8).

<sup>40</sup>Erlingsson, U., V. Pihur, and A. Korolova. (2014). "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response." *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*, 1054–1067.

<sup>41</sup> Ding, B., J. Kulkarni, and S. Yekhanin. (2017). "Collecting Telemetry Data Privately." *Advances in Neural Information Processing Systems* 30.

<sup>42</sup> Near, J. (2018) "Differential Privacy at Scale: Uber and Berkeley Collaboration," *Enigma 2018* (January) USENIX Assoc. <https://www.usenix.org/node/208168>.

<sup>43</sup> Bowen, C. et al. (2020) "A Synthetic Supplemental Public-Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications," (July) Tax Policy Center, The Brookings and Urban Institutes. [https://www.urban.org/sites/default/files/publication/102547/a-synthetic-supplemental-public-use-file-of-low-income-information-return-data\\_2.pdf](https://www.urban.org/sites/default/files/publication/102547/a-synthetic-supplemental-public-use-file-of-low-income-information-return-data_2.pdf).

<sup>44</sup> Australian Bureau of Statistics, (2019) "Protecting the Confidentiality of Providers," January 2019, *1504.0 - Methodological News*, <https://www.abs.gov.au/ausstats/abs@.nsf/Previousproducts/1504.0Main%20Features9999Jan%202019?opendocument&tabname=Summary&prodno=1504.0&issue=Jan%202019&num=&view=>, accessed on March 31, 2021.

Statistics in the United Kingdom,<sup>45</sup> and Statistics Canada<sup>46</sup> explicitly recognize the threat from combining multiple statistical tabulations to re-identify respondent information and recommend output noise infusion systems, including differential privacy.

46. Faced with the alarming results of the simulated reconstruction attack, which indicated that the established swapping mechanism resulted in far less disclosure protection than it was intended to provide, and considering the available alternatives, the Census Bureau's Data Stewardship Executive Policy Committee (DSEP)<sup>47</sup> determined that the Census Bureau should proceed with the deployment and testing of differential privacy for use in the 2020 Census given its obligations to produce high quality statistics from the decennial census while also protecting the confidentiality of respondents' census records under 13 U.S. Code §§ 8(b) & 9.<sup>48</sup>

---

<sup>45</sup> United Kingdom Office for National Statistics, (2021) "Policy on Protecting Confidentiality in Tables of Birth and Death Statistics," <https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyonprotectingconfidentialityintablesofbirthanddeathstatistics#annex-a-understanding-the-legal-and-policy-framework>, accessed on March 31, 2021.

<sup>46</sup> Statistics Canada, (2021) "A Brief Survey of Privacy Preserving Technologies," March 2021, *Data Science Network for the Federal Public Service*, <https://www.statcan.gc.ca/eng/data-science/network/privacy-preserving>, accessed on March 31, 2021.

<sup>47</sup> The Data Stewardship Executive Policy Committee (DSEP) is a committee chaired by the Deputy Director/Chief Operating Officer and composed of career senior executives with expertise in confidentiality practice, the uses of Census Bureau data, and policy. DSEP is the parent organization for the Disclosure Review Board (DRB), which reviews and approves individual data releases to ensure that no confidential data is released.

<sup>48</sup> On May 10-11, 2017 DSEP decided that "any request for disclosure avoidance of proposed publications for the 2020 Census be routed to the 2020 DAS team before going to the DRB" meaning that all 2020 Census publications would be subject to differential privacy. See Appendices E and F. On February 15, 2018 DSEP suspended publication of "all proposed tables in Summary File 1 and Summary File 2 for the 2020 Census at the block, block-group, tract, and county level except for the PL94-171 tables, as announced in Federal Register Notice 170824806-7806-01..." acknowledging that "...these data in many

47. The best disclosure avoidance option that offers a solution capable of addressing the new risks of reconstruction-abetted re-identification attacks, while preserving the fitness-for-use of the resulting data for the important governmental and societal uses of census data, is differential privacy. I have summarized here what I consider to be the most important reasons that the Census Bureau decided to adopt differential privacy.
48. **Disclosure avoidance must be proactive.** The fundamental objective of disclosure avoidance protections is to proactively prevent disclosures. Just like corporations are not expected to wait until they have suffered a major data breach before upgrading their IT security systems to protect against known threats, statistical agencies should not wait until they suffer a confirmed breach before improving their disclosure avoidance protections to account for known threats. The expectation, for both IT security and disclosure avoidance, is to remain vigilant about emerging threats and risks, and to take appropriate action *before* those risks lead to a breach.
49. **The privacy risk landscape has fundamentally changed since 2010.** Traditional methods of assessing disclosure risk rely on knowing what tools and resources an attacker might leverage to undermine confidentiality protections. These tools, however, are ever evolving. Over the last decade, technological advances have made powerful cloud computing environments, with sophisticated optimization algorithms

---

cases were accurate to a level that was not supported by the actual uses of those data, and such an approach is simply untenable in a formally private system.” DSEP further decided that “SF1 and SF2 will be rebuilt based on use cases.” See Appendix G. In parallel with these decisions by DSEP, the disclosure risks identified by the preliminary results of the simulated reconstruction attack also led to this issue being added to the Census Bureau’s risk management portfolio. On April 17, 2017 the risk of reconstruction attacks was proposed for inclusion in the Research and Methodology Directorate’s risk registry. On September 12, 2017 it was escalated and included on the Enterprise-level Risk register. Finally, on January 30, 2018, it was further escalated to the Enterprise-level Issue register, with the development and use of the 2020 Census Disclosure Avoidance System as an identified resolution action to be taken. .



capable of performing large-scale attacks, cheap and easily available. While these tools were not yet a viable attack model in 2010, they certainly represent a credible threat today.<sup>49</sup>

50. **Internal research has conclusively proven the fundamental vulnerabilities of the 2010 swapping methodology.** The Census Bureau has performed extensive empirical analysis of the disclosure risk inherent to the 2010 Census swapping methodology as detailed in Appendix B. No technique can produce usable data with absolutely zero risk of re-identification, but the re-identification rates from our internal experiments on the 2010 Census swapping methodology are orders of magnitude higher than what they were intended to be. The privacy threat landscape has evolved over the last decade and compels the Census Bureau to adapt its protections accordingly.

51. **The Census Bureau determined that differential privacy was the only method that could adequately protect the data while preserving the value of census data products.** When our internal research demonstrated the vulnerabilities of the swapping mechanism used for the 2010 Census, we considered a range of options for the 2020 Census. The three leading options were differential privacy, an enhanced version of data swapping, and a return to whole-table suppression. But to achieve the necessary level of privacy protection, both enhanced data swapping and suppression had severely deleterious effects on data quality and availability. With its enhanced privacy protections and precision control over the tuning of privacy/accuracy tradeoff, the Census Bureau determined that differential privacy was the only viable solution for the 2020 Census.

---

<sup>49</sup> DSEP drew this conclusion from the simulated reconstruction-abetted re-identification attack in Appendix B. The Office of National Statistics reached the same conclusion in its 2018 “Privacy and data confidentiality methods: a Data and Analysis Method Review (DAMR)” at [Privacy and data confidentiality methods: a Data and Analysis Method Review \(DAMR\) – GSS \(civilservice.gov.uk\)](#) (cited on April 10, 2021).

**52. Differential privacy can be fine-tuned to strike a balance between privacy and accuracy.** DSEP made the preliminary decision to pursue differential privacy on May 10-11, 2017. Since that decision was announced, the Census Bureau has worked extensively with our advisory committees, federal agency partners, American Indian and Alaska Native tribal leaders, the Committee on National Statistics, professional associations, data user groups, and many others at the national, state, and local levels to understand how they use decennial census data and to ensure that our implementation of differential privacy will preserve the value of the decennial census as a national resource. The Census also released sets of demonstrative data to allow the public and end-users to provide feedback that allowed us to fine-tune and tweak how we will ultimately implement differential privacy.<sup>50</sup>

**53. The need to modernize our privacy protections has been confirmed by external experts.** The Census Bureau's ongoing partnerships with scientific and academic experts from around the country helped us conduct the internal evaluation of the disclosure risk of the 2010 Census swapping methodology and confirmed the need to modernize our privacy protections. To supplement this ongoing work and to get external expert confirmation of the conclusions that we have drawn from it, the Census Bureau also commissioned an independent expert review by JASON, an independent group of elite scientists that advise the federal government on science and technology. The JASON report confirmed our findings regarding the re-identification risk inherent to the 2010 Census swapping methodology.<sup>51</sup>

---

<sup>50</sup> U.S. Census Bureau "Developing the DAS: Demonstration Data and Progress Metrics" <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-development.html>.

<sup>51</sup> JASON (2020). "Formal Privacy Methods for the 2020 Census" JASON Report JSR-19-2F. <https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/privacy-methods-2020-census.pdf>.

54. **Differential Privacy can produce highly accurate data.** One key benefit of differential privacy is the ability to fine-tune privacy and accuracy. The next iteration of demonstration data will establish that differential privacy protections can produce extremely accurate redistricting data. While the full April 2021 Demonstration Data Product<sup>52</sup> and supporting metrics will be released by April 30, 2021, I can provide a high-level summary of key metrics:<sup>53</sup>

- Total populations for counties have an average error of +/- 5 persons (reflecting a mean absolute percent error of 0.04% of the counties' population) as noise from differential privacy.<sup>54</sup> This is extremely accurate considering that if we simulate the errors in census counts as estimates of the true population, then the average county-level estimation uncertainty of the census is +/- 960 persons (averaging 1.6% of the county census counts).<sup>55</sup>

---

<sup>52</sup> The April 2021 demonstration data uses a global privacy-loss budget of 10.3 with a very substantial proportion allocated to detailed race and ethnicity statistics at the block and block group levels.

<sup>53</sup> Statistics for the April 2021 Demonstration Data Product are preliminary, based on the internal research version. The production version will be used for the detailed summary statistics when they are posted on census.gov.

<sup>54</sup> The statistics are the mean absolute error and the mean absolute percentage error in county population comparing the April 2021 Demonstration Data Product to the data released in the 2010 Summary File 1.

<sup>55</sup> The inherent error in the census counts as estimates of the true population can be simulated using data-defined person and correct-enumeration rates from coverage measurement estimates, in this case from the most recent decennial census in 2010. (See Mule, T. "2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States", Report G-10, g01.pdf (census.gov). Table 3, in particular.) An alternative modeling perspective simulates the natural variation of census population estimates using the natural variation in census estimates due to erroneous enumerations and other sources of error inherent in the Census. For county populations

- At the block level the differentially private data have an average population error of +/- 3 persons, which includes both housing unit and group quarters populations. Compare that with the simulated error inherent in the census which puts the average error uncertainty of block population counts at +/- 6 people.<sup>56</sup>

55. **The April 2021 demonstration data show no meaningful bias in the statistics for racial and ethnic minorities** even in very small population geographies like Federal American Indian Reservations. The data permit assessment of the largest OMB-designated race and ethnicity group in each geography – the classification used by the Department of Justice for Voting Rights Act scrutiny – with a precision of 99.5% confidence in variations of +/- 5 percentage points for off-spine geographies as small as 500 persons, approximately the minimum voting district size in the redistricting plans that the Department of Justice provided as examples.

56. **The accuracy of differential privacy increases at higher levels of geography, even for arbitrary geographic areas like Congressional and legislative districts.** The Census Bureau designed its implementation of differential privacy to increase accuracy

---

this natural variation is about +/- 120 persons (0.3% of population), also based on coverage data from the 2010 Census. As with all simulation estimates, there is sensitivity to the assumptions. The reported statistics are the mean absolute error and the mean absolute percentage error. Differentially private statistics include both the housing unit and group quarters populations. Simulations exclude the group quarters population because there are no coverage estimates for that group.

<sup>56</sup> The simulation of the natural variation of census block-level populations is +/- 1.5 persons, which excludes the group quarters population because there are no coverage estimates for that group. As with all simulation estimates, there is sensitivity to the assumptions. The reported statistics are the mean absolute errors. Mean absolute percentage errors are not useful statistics for block populations because more than 2,000,000 blocks with positive housing units have populations between 0 and 9. Differentially private statistics include both the housing unit and group quarters populations. Simulations exclude the group quarters population because there are no coverage estimates for that group.

as blocks are aggregated into larger geographic areas like neighborhoods, voting districts, towns, and other places. Rather than infusing noise at the block level and aggregating upwards, which would cause error to compound at larger geographic levels, the Disclosure Avoidance System's TopDown Algorithm (TDA) takes the opposite approach. Starting at the national level, the algorithm establishes very precise (but still privacy-protected) tabulations for all characteristics at the national level, then works its way down the geographic hierarchy, ensuring that all of the geographic entities at each level (e.g., the Census tracts within a county) add up precisely to the established characteristics of the level above (e.g., the county). This approach limits the distortions that can arise from noise infusion and ensures the reliability of statistics as the underlying size of the population increases. Plaintiffs argue that "the November 2020 demonstration data also skewed the 2010 tabulations enough to create a population deviation in Alabama's Congressional districts on a level that courts have found in other contexts to violate voters' equal population rights," with districts losing up to 73 individuals or gaining 206 individuals over reported values. While this may have been true for the November 2020 Demonstration Data Product, this is not true for the Demonstration Data Product that will be produced by the end of April. In the April 2021 Demonstration Data Product, Congressional districts as drawn in 2010 have a mean absolute percentage error of 0.06%. If the Congressional districts had been drawn using the April 2021 Demonstration Data Product, their statistical composition for the purposes of Voting Rights Act scrutiny would not be affected. Even for state legislative districts, which had average sizes of 159,000 (upper chambers) and 64,000 (lower chamber), the mean absolute percentage errors are 0.09% (upper chambers) and 0.16% (lower chambers), respectively. Such errors are trivial and imply that the difference between districts drawn from the April 2021 Demonstration Data Product and those drawn from the original 2010 P.L. 94-171 Redistricting Data Summary File would be statistically and practically imperceptible. *Most importantly*

*for the redistricting use case, the TDA, when properly tuned, ensures that redistricters can remain confident in the accuracy of the population counts and demographic characteristics of the voting districts they draw, despite the noise in the individual building blocks.*

#### **IMPLEMENTING DIFFERENTIAL PRIVACY FOR THE 2020 CENSUS**

57. Census announced that it planned to use Differential Privacy for the 2020 Census in a few different venues: (1) August 3, 2018, 2020 Census Program Management Review; (2) December 6, 2018, Census Scientific Advisory Committee Meeting; and (3) May 2, 2019, Census National Advisory Committee meeting.
58. The Bureau has engaged in a years-long campaign to educate the user community and solicit their views about how differential privacy should be implemented. Census Bureau staff have made hundreds of public presentations, held dozens of webinars, held formal consultations with American Indian and Alaska Native tribal leaders, created an extensive website with plain English blog posts, and conducted regular outreach with dozens of stakeholder groups. We have made presentations to our scientific advisory committees and provided substantial information to oversight entities such as the Government Accountability Office and the Office of the Inspector General.
59. Part of the Bureau's effort to inform the public and solicit feedback involved releasing a series of Demonstration Data Products. There are many different ways to implement differentially private disclosure avoidance mechanisms, and the design and parameters of these mechanisms can substantially impact the fitness-for-use of the resulting data. The Census Bureau's TopDown Algorithm (TDA) was specifically designed to address the reconstruction-abetted re-identification vulnerability risks, while allowing the Bureau to tune the accuracy of the statistics to ensure fitness-for-use.

60. To date, the Census Bureau has released four sets of Demonstration Data Products (in October 2019, May 2020, September 2020, and November 2020). The Census Bureau has received substantial, actionable feedback after each release that has contributed to the system’s design and optimization.
61. All four of these demonstration products used a lower privacy-loss budget than we anticipate using for the final 2020 Census data – that is, these demonstration data were purposefully “tuned” to privacy and not “tuned” for producing highly accurate re-districting data. We held the privacy-loss budget roughly the same across these four releases to allow us to compare effects of incremental improvements in the system. After each release, these demonstration files enabled data users to help the Census Bureau identify areas where the algorithm needed to be tuned to meet their specific use cases. While the Census Bureau has not yet set the final privacy-loss budget, we have been clear that all the demonstration data released to date have used a lower privacy-loss budget (more privacy, less accuracy) than will be selected for the final production run of the redistricting data.<sup>57</sup>
62. This degree of transparency into the design and implementation of a disclosure avoidance methodology is unprecedented in the federal government. The Census Bureau has submitted its differential privacy mechanisms, programming code, and system architecture to thorough outside peer review. We have also committed to publicly releasing the entire production code base and full suite of implementation settings and parameters. Many traditional disclosure avoidance methods, most notably swapping techniques, must be implemented in a “black box.” Implementation parameters for these legacy disclosure avoidance methods, especially swapping rates, are often

---

<sup>57</sup> Most recently on February 23, 2021 in [The Road Ahead: Upcoming Disclosure Avoidance System Milestones \(govdelivery.com\)](https://www.govdelivery.com).

some of the most tightly guarded secrets that the Census Bureau protects. But differential privacy does not rely on the obfuscation of its implementation as a means of protecting the data. The Census Bureau's transparency will allow any interested party to review exactly how the algorithm was applied to the 2020 Census data, and to independently verify that there was no improper or partisan manipulation of the data.

**INVARIANTS ARE NOT REQUIRED FOR ACCURACY.**

63. Invariants – or data held constant when applying differential privacy – introduce privacy risks and are not necessary to ensure accuracy. Invariants were not well understood either theoretically or empirically in 2016 when the Census Bureau began its research on differential privacy for decennial census data, but we now understand that invariants defeat the privacy protections and must be limited in order to protect the integrity of the system as a whole. Unlike traditional approaches to disclosure avoidance, differentially private noise infusion offers quantifiable and provable privacy guarantees. These guarantees, reflected in the global privacy-loss budget and its allocation to each statistic, serve as a promise to data subjects that there is an inviolable upper bound to the risk that an attacker can learn or infer something about those data subjects through publicly released data products. While that upper bound is ultimately a policy decision, and may be low or high depending on the balancing of the countervailing obligations to produce accurate data and to protect respondent confidentiality, the level of the global privacy-loss budget is central to the ability of the approach to protect the data. Invariants are, by their very nature, the equivalent of assigning infinite privacy-loss budget to particular statistics, which fundamentally violates the central promise of differentially private solutions to controlling disclosure risk. By excluding the accuracy of invariant data elements from the control of the privacy-loss budget, invariants exclude the disclosure risk and potential inferences that can be drawn from those data elements from the formal privacy guarantees. Thus,



instead of being able to promise data subjects that the publication of data products will limit an attacker to being able to infer, at most, a certain amount about them (with that amount being determined by the size of the privacy-loss budget and its allocation to each characteristic), the inclusion of one or more invariants fundamentally excludes attacker inferences about the invariant characteristic(s) from the very nature of that promise. The qualifications and exclusions to the privacy guarantee weaken the strength of the approach and make communicating the resulting level of protection substantially more difficult. This is the reason that DSEP removed the block-level invariant on population and voting-age population. Below the state level, DSEP only authorized block-level invariants that were necessary to conduct the field operations of the 2020 Census: housing unit address counts, and occupied group quarters address counts and types. As noted above, if the block population is reported with some random fluctuation around the confidential value, then only by chance will the block identifier be correct in any potential reconstructed microdata. Compound this effect over 8,000,000 blocks and the number of feasible reconstructions explodes exponentially. This is what provides the protection against re-identification from the reconstructed data.

64. Invariants are not required to improve the accuracy of any statistic processed by differential privacy. Assigning sufficiently high (but not infinite) privacy-loss budget to any statistic can ensure perfect accuracy for that statistic while still allowing the resulting privacy-loss to be communicated in the privacy guarantee. For example, the state-level population of the American Indian and Alaska Native tribal areas has been given sufficient privacy-loss budget to ensure that those populations are presented accurate to the number of persons in the units column; the mean absolute error is 1 person, essentially invariant and the same precision as the state populations themselves. But this solution still requires balancing accuracy and privacy-loss overall. All characteristics cannot have large privacy-loss budget allocations at every geographic

level. If they did, the published tables would be exact images of the confidential data and subject to the same vulnerability as the 2010 Census.

65. The forthcoming April 2021 Demonstration Data Product illustrates this tradeoff. These new demonstration data use a global privacy-loss budget for persons of 10.3, which is much larger than the 4.0 budget used in the earlier releases but is still allocated in a manner that provides a level of protection for every census record and every published characteristic. The April 2021 demonstration data also fully satisfy a tightly specified set of accuracy criteria specialized to the redistricting use case. Specifically, populations, voting-age populations, and the proportion of the largest OMB-designated race and ethnicity groups are all reliable for redistricting and Voting Rights Act scrutiny in arbitrary contiguous block aggregates for both on-spine and off-spine political and legal entities. Because new districts cannot be drawn before the 2020 P.L. 94-171 Redistricting Data Summary File is released, counties, block groups, minor civil divisions, incorporated places, and Census-designated places were all used as on- and off-spine geographic entities for tuning purposes.
66. In the April 2021 Demonstration Data Product, all the targeted small population statistics for race and ethnic groups are far more accurate than in previous demonstration data products, even though no additional invariants were used. The gain in accuracy is entirely due to dedicating more of the privacy-loss budget to the block- and block group-level statistical tables and carefully specifying the differentially private measurements to target the OMB-designated race and ethnicity groups. Biases in the tribal areas' race and ethnicity data were also greatly reduced.
67. The Census Bureau has received substantial feedback from our data user community highlighting distortions that were present in the early versions of our demonstration data, particularly in the version released in October 2019. Based on that feedback, the Census Bureau has identified and corrected the algorithmic sources of those distortions. As these measures of accuracy and bias show, any residual impact of the types

of systematic bias observed in the early demonstration data will be negligible and well within the normal variance and total error typical for a census.

#### **PROCESS AND TIMELINE MOVING FORWARD**

68. The operational delays caused by the global COVID-19 pandemic, and the resulting processing schedule changes for production of the redistricting data product shifted the milestone dates for all the systems necessary to produce the data. While the 2020 Census Disclosure Avoidance System is fully operational, and has already passed the Test Readiness Review (TRR) and Production Readiness Review (PRR) milestones on schedule, we have taken advantage of the additional time before the May 20, 2021 Operational Readiness Review (ORR) to perform additional optimization and testing of the system, and to engage in another round of data user evaluation and feedback.
69. The Census Bureau will release another demonstration product by April 30, 2021 using a higher privacy-loss budget (more accuracy) that better approximates the final privacy-loss budget that will likely be selected for the redistricting data product. These new demonstration data will also reflect system design changes that have been made since the last demonstration data release, along with tuning and optimization of the system that have been done specifically to prioritize population count accuracy and the ability to identify majority-minority districts.<sup>58</sup> The new release will give users yet another opportunity to let the Census know specifically where the data are (or are not yet) sufficiently accurate to meet their requirements.
70. On March 25, 2021, DSEP approved the privacy-loss budget to be used for the next demonstration product. This privacy-loss budget reflects empirical analysis of over

---

<sup>58</sup> Users will be able to see the difference between algorithmic improvements and greater privacy-loss budget. At the same time as the main April 2021 Demonstration Data Product is released, the Census Bureau will also release demonstration data using exactly the same software implementation but setting the global privacy-loss budget to 4.0 for persons, as it was in the four previous demonstration data products.

600 full-scale runs of the Disclosure Avoidance System using 2010 Census data. The Census evaluated these experimental runs using accuracy and fitness-for-use criteria for the redistricting use case informed by the extensive feedback we have received from the redistricting community and the Civil Rights Division at the U.S. Department of Justice. Based on this feedback, the privacy-loss budget for the final demonstration product is set to ensure the accuracy of racial demographics for voting districts as small as 500 individuals. With this tuning, the proportion of the largest racial group within even those small state/local voting districts of 500 individuals will be accurate to within five percentage points of the enumerated value at least 95% of the time. As voting district population size increases to any sort of reasonably anticipated legislative district, the error will be miniscule. For example, Congressional and state legislature districts will have significantly higher accuracy for population counts and voting age population counts.

71. Following the release of the new demonstration data, data users and stakeholders will have about a month to submit additional feedback on their analysis and assessment of these data, before DSEP, in early June 2021, sets the privacy-loss budget and system parameters for the production run of the redistricting data product.
72. The production run for creating the Microdata Detail File (the internal name for the file that contains the privacy-protected data) is scheduled to occur between June 26 and July 18, 2021. This roughly three-week period is similar to the period required to implement disclosure avoidance in prior censuses and is not the cause of the delay in the delivery of the redistricting data.
73. As discussed in more detail below, any court-ordered change in the Census Bureau's implementation of disclosure avoidance would add significant time to this schedule.

**BRYAN AND BARBER DECLARATIONS**

74. Although I cannot set out all my observations and disagreements with the declarations of Dr. Michael Barber and Mr. Thomas Bryan in this declaration, I want to identify some key areas of dispute.

75. Dr. Barber's expert report does not adequately account for the fact that the Census Bureau's demonstration data products had a privacy-loss budget significantly lower than the expected budget that will be set for the 2020 Census. As I explained above, we purposefully set the budget lower than ones most likely to be finally chosen (set to favor privacy over accuracy), so that we could isolate the distortions and demonstrate the effectiveness of various methodological modifications. One cannot draw conclusions about the accuracy of the data the Census Bureau will release for the 2020 Census based on these demonstration products.

76. Dr. Barber is premature in drawing conclusions about the accuracy of the 2020 redistricting data before the Census Bureau has set a final privacy-loss budget, and he is further incorrect in opining on the accuracy of differential privacy without considering the relative error of alternatives. Dr. Barber focuses most of his report on the possible quality concerns of differentially private 2020 Census data releases with no attention to (1) the demonstrated privacy risks of a 2020 Census protected by legacy methods and (2) the accuracy of alternatives to differential privacy including enhanced swapping or suppression. As I show in this declaration, all disclosure avoidance systems trade-off accuracy for confidentiality protection. They must be compared to each other. Releasing the redistricting data without disclosure avoidance procedures – tabulating the Census Edited File directly – is not an option and was not done for the 1990, 2000, or 2010 Censuses.

77. Dr. Barber relies on external studies that draw incorrect conclusions and use early demonstration data products. In his declaration, Dr. Barber quotes Santos-Lozada, et al. (2020) on page 14 by saying that “[i]nfusing noise in the data, in comparison to the

current disclosure avoidance system, will produce inaccurate patterns of demographic change with higher levels of error found in the calculations for non-Hispanic blacks and Hispanics. At the same time, these counts are bound to impact post-2020 districting for both federal and state elections, as well as evaluations of that redistricting. . . . [T]hese changes in population counts will affect understandings of health disparities in the nation, leading to overestimates of population-level health metrics of minority populations in smaller areas and underestimates of mortality levels in more populated ones.” The Santos-Lozada et al. paper uses the October 2019 Demonstration Data Product. Therefore, its conclusions are only applicable to the state of the algorithms and the overall privacy-loss budget used for that early release. Those were neither the final algorithms nor the final privacy-loss budget. I informed the editors of the Proceedings of the National Academy of Sciences of these defects during the peer-review process. I strongly recommended that the word “will” in the title be changed to “may” for these reasons. There is nothing statistically incorrect in the paper except for the general failure of these demographers to account for estimation error due to disclosure avoidance when doing their statistical analyses as I have noted in my own scholarly work<sup>59</sup> and other statisticians and computer scientists have also noted.<sup>60</sup> The fatal error in the Santos-Lozada et al. paper is drawing conclusions from preliminary data generated by an obsolete version of the 2020 Census DAS using obsolete settings for the privacy-loss budget and its allocation. Those conclusions are wrong and so, by extension, are those of Dr. Barber.

---

<sup>59</sup> Abowd, John M. and Ian Schmutte “Economic Analysis and Statistical Disclosure Limitation” *Brookings Panel on Economic Activity* (Spring 2015): 221-267. [[download article and discussion, open access](#)] [[download preprint](#)].

<sup>60</sup> Wasserman L. and S. Zhou “A Statistical Framework for Differential Privacy,” *Journal of the American Statistical Association*, Vol. 105, No. 489 (2010):375-389, DOI: [10.1198/jasa.2009.tm08651](https://doi.org/10.1198/jasa.2009.tm08651).

78. Dr. Barber's conclusions do not take into account that if the Census Bureau were forced to hold the number of people in housing units invariant at the block level, that would, in turn, require adding more noise and error to the demographic characteristics of those individuals in an effort to offset what amounts to assigning block-level populations an infinite privacy-loss budget. As I show in my declaration, doing so is unnecessary and harmful to both accuracy and confidentiality protection. The correct procedure is to set accuracy targets for meaningful aggregations then tune the disclosure avoidance procedures to meet them. This procedure is transparent when using differential privacy, but it was also done for the 2010 swapping system albeit in memos that are also protected by 13 U.S. Code §§ 8(b) & 9.
79. Furthermore, Dr. Barber's work draws incorrect conclusions about biases in rural areas and for specific small populations. In his declaration, Dr. Barber states on page 13 that "[p]laces with fewer people (rural locations) and areas with smaller, distinctive populations (minority communities) are more likely to be impacted since these are the places where identification is more concerning, and the application of statistical noise is more likely to have a larger impact on the summary statistics derived from the altered data." He concludes on pages 13 and 14 that "...the process of differential privacy is not applied equally across the entire population. Places with fewer people (rural locations) and areas with smaller, distinctive populations (minority communities) are more likely to be impacted since these are the places where identification is more concerning, and the application of statistical noise is more likely to have a larger impact on the summary statistics derived from the altered data." This conclusion is incorrect. His analysis should say that the privacy-loss of the respondents in these small areas is being treated equally and identically to the privacy-loss of the respondents in large population areas; that is, every single respondent gets the full privacy protection afforded by the DAS—unlike the 2010 system, which only tried to protect certain households. To properly compare urban/rural statistics before and after the

application of disclosure avoidance, regardless of the system, the full algorithm assigning rural/urban status must be used on both the privacy-protected and confidential data. Dr. Barber has not done this.

80. Dr. Barber's work makes incorrect assertions about the non-negativity constraint. In his declaration, Dr. Barber cites Riper, Kugler, and Ruggles (2020) on page 13 stating that "[t]he non-negativity constraint requires that every cell in the final detailed histogram be non-negative. As described above, many of the cells in the noisy household histograms will be negative, especially for geographic units with smaller numbers of households. Returning these cells to zero effectively adds households to these small places, resulting in positive bias." This point is not an accurate description of how non-negativity is being handled in the post-processing of the noisy histogram. The analysis should say that negative values are not simply being returned to zero, but that all blocks with housing units are used to estimate the population counts subject to a non-negativity constraint on the solutions. That is, negative values are not "[r]eturning to zero," the entire 2,016 element matrix (for the redistricting data) is smoothed to a consistent, non-negative matrix for each of the 8,000,000 blocks, 275,000 block groups, 75,000 tracts, 3,143 counties, 51 states (including DC), and the U.S. simultaneously.<sup>61</sup> At the block-level, there are expected to be an average of only 40 people represented across the 2,016 cells. This is the inherent sparsity that any disclosure avoidance system must address. Dr. Barber claims on page 13 that "[t]he combination of the non-negativity constraint and population invariants consistently leads to bias increasing counts of small subgroups and small geographic units and decreasing counts of larger subgroups and geographic units." While the statement is correct in

---

<sup>61</sup> The matrix is 2,016 elements rather than 252 because there are eight elements in the Group Quarters Table P5 (seven group quarter types and "not a group quarters") that also interact with the other categories. The number of geographic entities at each level is based on approximate values for 2020 tabulation geographies.



principle, the magnitudes shown in his report are not representative of the final re-districting data product. At the levels of privacy-loss budget used for the forthcoming April 2021 Demonstration Data Product, the consequences of the non-negativity constraint were tightly controlled for population areas of at least 500 total persons. The remaining variation in block-level statistics, including small biases, is required to protect locational privacy and deliver consistent data. It is well within the inherent variability of block-level census data, as shown in my declaration.

81. Dr. Barber argues that the amount of error observed in the demonstration files indicates that differential privacy cannot produce data sufficient for important use cases. Mr. Barber's focus on the percentage of blocks in the demonstration data that differ at all from the official 2010 Census data (even if that difference represents the addition or subtraction of a single individual from the block) ignores two important points. First, the entire objective of our implementation of differential privacy is to infuse sufficient noise in block-level data to protect against reconstruction-abetted re-identification attacks while ensuring that when those blocks are aggregated into larger geographies of interest (voting districts, towns, etc.) those relative errors diminish and the accuracy of the tabulations improves. Second, the overall accuracy of the data is a direct consequence of the global privacy-loss budget selected and how it is allocated. The demonstration data used by both Dr. Barber and Mr. Bryan for their analyses, which use a substantially lower privacy-loss budgets than will be used for the final 2020 Census data products, can therefore be expected to be substantially "noisier" than the final data will be. Examples of noise levels in the April 2021 Demonstration Data Product provided in my report and verifiable when those data are released later this month confirm my claims.
82. Mr. Bryan assesses the accuracy of the four Demonstration Data Products (October 2019, May 2020, September 2020 and November 2020) using the percent of blocks with any change at all (pp. 9-13) or percentage errors (pp. 16-19). Both sets of analyses are

based on obsolete versions of the DAS, but they also make serious errors that will still be salient when he uses the April 2021 Demonstration Data Product. The DAS was designed to control the error in counts, not percentages. The basic tables in the P.L. 94-171 Redistricting Data Summary File are counts of resident persons living in specific geographies who have features chosen from the following taxonomy {any age, voting age}, {Hispanic/Latino, not Hispanic/Latino}, and any combination of {Afro-American/Black, American Indian/Alaska Native, Asian, Native Hawai'ian/Pacific Islander, White, Some other race} except "none." The specific aggregate geographies available in the data product are all built from census blocks, but it is the counts of persons in those aggregate geographies, including voting districts, not the block counts themselves that must be accurate enough to be fit for redistricting. Block-level errors, whether in counts or percentages, are irrelevant except to the extent that they are not controlled in larger-population geographies. In 2010, the average population in a block was 28 and the average population in an occupied block was 49. Any block-level variation in one of the 2,016 cells of the redistricting data for total populations this small is going to appear as a "large" percentage error. Indeed, most of those statistics have a base of zero, making percentage variation undefined and meaningless. The DAS must introduce noise into the block-level data to achieve any confidentiality protection at all. This statement is also true for the systems that were used in the 1970 to 2010 Census. The noise from suppression (1970, 1980) is counts that are simply not reported at the block level. The noise from blank and impute (1990) is due to the imputation modeling. The noise from swapping (2000, 2010) is due the exchange of geographic identifiers across blocks. All confidentiality protection applied to block-level redistricting data produces errors of the sort described by Mr. Bryan. Furthermore, many of the supposed DAS errors in Mr. Bryan's analysis cancel out when blocks are aggregated into larger-population geographies like block groups, census tracts, towns, counties, and congressional districts. This is not an accident; it is a carefully

designed feature of the DAS. The tabulation of the protected microdata might miss a person in one block, but have an “excess” person in the neighboring block for a particular characteristic. Because the DAS uses direct measurements from the U.S. all the way down to the block to estimate the counts at every level of geography, whether on- or off-spine, they are all much more accurate than any of the block estimates that comprise them. This is easy to see in any balanced summary of the accuracy of the DAS. Counties and places have far smaller percentage errors than the average percentage error of the blocks that compose them.

#### CLARIFYING STATEMENT QUOTED IN COMPLAINT

83. Plaintiffs assert, quoting an article in 2018 by the demographer Steven Ruggles and others, that I claimed that database reconstruction does not pose a significant re-identification threat. I made the statement that plaintiffs reference indirectly at the December 14, 2018 meeting of the Federal Economic Statistics Advisory Committee (FESAC) in my own presentation.<sup>62</sup> Dr. Ruggles was on the FESAC program in the same session. I made the remarks in December 2018 as a report on ongoing research.<sup>63</sup> At the February 16, 2019 session of the American Association for the Advancement of Science (AAAS), I retracted my tentative conclusion about re-identification based on additional research reported there. The full text and presentation of the AAAS session are attached as Appendices H and I.<sup>64</sup> To be clear, the Census Bureau’s simulated recon-

---

<sup>62</sup> Federal Economic Statistics Advisory Committee program: [FESAC Meeting Agenda December 2018 \(bea.gov\)](#).

<sup>63</sup> My remarks at the December 18, 2018 FESAC: [Microsoft PowerPoint - Abowd Presentation \(bea.gov\)](#).

<sup>64</sup> AAAS materials for the February 16, 2019 session area also here: <https://blogs.cornell.edu/abowd/files/2019/04/2019-02-16-Abowd-AAAS-Talk-Saturday-330-500->

struction attack on the 2010 Census data described in this declaration and in the accompanying appendix materials shows there is a significant re-identification risk. However, the Census Bureau's Data Stewardship Executive Policy Committee (DSEP) acted to adopt differential privacy as soon as that research showed that an accurate microdata reconstruction was feasible. It did not require, nor should it have required, the subsequent demonstration that those reconstructed microdata permit between 52 and 179 million correct re-identifications from the 2010 Census. The reconstructed microdata fail the *2010 Census* microdata disclosure avoidance requirements—the requirements that were in place for that census—because they contain geographic identifiers (the block code) that relate to a minimum population of one rather than the 100,000 person minimum population that contemporary standards required. The reconstructed microdata also did not impose any of the minimum population thresholds required of the tabulation variables, especially age.<sup>65</sup> These requirements were already in place because it is well understood at the Census Bureau and in the official statistics community worldwide that geographic identifiers for low-population areas, sex, and exact age in microdata files are a major disclosure risk especially in population censuses.

#### **IMPACT OF ANY COURT RULING BARRING USE OF DIFFERENTIAL PRIVACY**

84. Were the Court to rule that the Census Bureau was precluded from using differential privacy for the 2020 Census P.L. 94-171 Redistricting Data Summary File, we would be faced with hard choices. The inevitable result would be significant delay in deliv-

---

[session-FINAL-as-delivered-2jr4lzb.pdf](#) and <https://blogs.cornell.edu/abowd/files/2019/04/2019-02-16-Abowd-AAAS-Slides-Saturday-330-500-session-FINAL-as-delivered-1iqsdg2.pdf>.

<sup>65</sup> McKenna (2019a).

ery of the already-delayed redistricting data and diminished accuracy. Either the Census Bureau would have to revert to using suppression (as was last used in the 1980 Census) or use enhanced swapping (as was used in the 1990 to 2010 Censuses, but at a much higher rate and with fewer invariants). Either choice would delay results and diminish accuracy.

85. The effect on the schedule for delivering redistricting data would be substantial. The Census Bureau cannot ascertain the length of the delay until it understands any parameters the Court might place on its choice of methodology, but under all scenarios the delay would be multiple months. This delay is unavoidable because the Census Bureau would need to develop and test new systems and software, then use them in production and subject the results to expert subject matter review prior to production of data. The Census Bureau has been developing the systems and software to use differential privacy for several years – the agency has spent millions of dollars purchasing cloud computer capacity and writing and tuning code. The systems and software are ready to go and await only final tuning and a decision on the privacy-loss budget.
86. Even if the agency was ordered to repeat exactly what was done in 2010 (despite the serious risks to privacy the Census has identified), we could not simply “flip a switch” and revert to the prior methodology. Instead, we would need to conduct the requisite software development and testing. The 2020 Census’s system architecture is completely different than that used in the 2010 Census, and it is thus not possible to simply “plug in” the disclosure-avoidance system used in 2010.
87. Not only would redistricting data be further delayed, but the resulting data would be less accurate. Both swapping and suppression are blunt instruments for privacy protection. Unlike differential privacy, neither can be effectively tuned to optimize for data accuracy. Knowing that the 2010 Census results were vulnerable to reconstruction, the Census Bureau cannot simply repeat the swapping protocols from the 2010

census, but rather would be forced to fashion appropriate levels of protection for either system. Using an appropriate level of protection for either suppression or swapping would produce far less accurate data than would differential privacy.

88. I would urge any court to be quite wary of opining on the suitability of particular methods for conducting disclosure avoidance, as these decisions are highly technical and can have unanticipated consequences. The only reason the Court knows so much about the proposed methods for the 2020 Census is that transparency does not undermine their confidentiality protections, which is not the case for either swapping or suppression. While we cannot predict the full impact of any change, there is a danger than any change would have cascading effects on data accuracy and privacy, making race and ethnicity data, along with age data, substantially less accurate. Any sort of change in the basic methodology would be minimally tested and would not have the benefit of any input from the user community.

89. In conclusion, it is my professional opinion that the Census Bureau's Data Stewardship Executive Policy Committee should be permitted to control the type and parameters of any disclosure avoidance system used for the 2020 Census, just as it did for the 2010 Census and just as its predecessor committees did for decennial censuses conducted since the passage of the Census Act (13 U.S. Code) in 1954.

I declare under penalty of perjury that the foregoing is true and correct.

DATED and SIGNED:

**JOHN ABOWD**

---

Digitally signed by JOHN ABOWD  
Date: 2021.04.13 08:45:14 -04'00'

John M. Abowd

Chief Scientist and Associate Director for Research and Methodology

United States Bureau of the Census

# John M. Abowd

[Home](#)   [Professional Information](#)   [Courses](#)   [Recent News](#)   [Special Materials](#)

## Professional Information

[Updated April 1, 2021]

### CONTACT INFORMATION

U.S. Census Bureau  
HQ 8H120 ATTN: Sara Sullivan  
4600 Silver Hill Road  
*Private delivery services (FedEx, UPS, etc.) physical location: Suitland, MD 20746*  
*USPS mail only: Washington, DC 20233*  
Voice: +1.301.763.5880  
Mobile: +1.202.591.0766  
Fax: +1.301.763.8360  
Executive assistant Sara Sullivan: +1.301.763.5116  
E-mail: [john.maron.abowd@census.gov](mailto:john.maron.abowd@census.gov)

ILR School  
*USPS mail only (send private delivery service items to the address above):*  
275 Ives Hall  
Cornell University  
Ithaca, New York 14853-3901  
Assistant: [LDI@cornell.edu](mailto:LDI@cornell.edu)  
E-mail: [john.abowd@cornell.edu](mailto:john.abowd@cornell.edu)

Webpage: <https://blogs.cornell.edu/abowd/> or <https://www.johnabowd.com>

Twitter: [@john\\_abowd](https://twitter.com/john_abowd) (opinions are my own)

Short biography in PDF format

### CURRENT POSITIONS

Chief Scientist and Associate Director for Research and Methodology, U. S. Census Bureau, **IPA** June 1, 2016 – March 27, 2020; Career Senior Executive Service March 29, 2020 –

Edmund Ezra Day Professor, Department of Economics, Cornell University, July 2011 – **currently on leave**

Director, Labor Dynamics Institute, Cornell University, October 2011 – **currently on leave**

Founding member and Professor of Information Science (by courtesy), Faculty of Computing and Information Science, July 2000 – **currently on leave**

Professor of Statistics and Data Science, September 2013 – **currently on leave**

Member of the Graduate Fields of Economics, Industrial and Labor Relations, Information Science, and Statistics

Search ...

SEARCH

#### INSTITUTIONS

[U.S. Census Bureau](#)

[Cornell Economics](#)

[Labor Dynamics Institute](#)

[NCRN node at Cornell](#)

[CISER](#)

#### OTHER INFORMATION

[Google Scholar](#)

[ORCID](#)

[RePEC/Ideas](#)

[SSRN](#)

Research Associate, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, Massachusetts 02138, September 1983 – (on leave while serving at the U.S. Census Bureau)

Research Affiliate, Centre de Recherche en Economie et Statistique/INSEE, 15, bd Gabriel Péri, 92245 Malakoff Cedex France, November 1997 –

Research Fellow, IZA (Institute for the Study of Labor), P.O. Box 7240 D-53072 Bonn, Germany, June 2002 –

Research Fellow, IAB (Institut für Arbeitsmarkt-und Berufsforschung), Dienstgebäude Weddigenstraße 20-22, 90478 Nürnberg, Germany, January 2013 –

President and Principal, ACES-Research, LLC, john@aces-research.com, July 2007 –

Editor, Journal of Privacy and Confidentiality Online journal

## PREVIOUS AND VISITING POSITIONS

Distinguished Senior Research Fellow, United States Census Bureau, September 1998 – May 2016

Associate Chair, Department of Economics, Cornell University, August 2015 – May 2016

Visiting Professor, Center for Labor Economics, University of California-Berkeley, August 2014 – July 2015

Director of Graduate Studies, Economics, July 2010 – June 2014

Professor of Economics and Econometrics, University of Notre Dame, January 2008 – May 2008.

Director, Cornell Institute for Social and Economic Research (CISER), July 1999 – December 2007

Associate Director, Cornell Theory Center (became Cornell University Center for Advanced Computing), September, 2006 – August 2007.

Professor of Labor Economics, Cornell University, January 1990 – October 2001.

Edmund Ezra Day Professor, School of Industrial and Labor Relations, November 2001 –

Associate Director, Cornell Institute for Social and Economic Research (CISER), July 1998 – June 1999.

Chair, Department of Labor Economics, Cornell University, September 1992 – June 1998.

Acting Director, CISER, January 1998-June 1998.

Professeur invité, Laboratoire de Microéconomie Appliquée-Theorie Et Applications en Microéconomie et macroéconomie (LAMIA-TEAM), Université de Paris-I (Panthéon-Sorbonne), May 1998.

Consultant, Centre de Recherche en Economie et Statistique (CREST), Institut National de la Statistique et des Etudes Economiques (INSEE), February 1997.

Professeur invité, ERMES (Equipe de Recherche sur les Marchés, l'Emploi et la Simulation) Université Panthéon-Assas (Paris II), October 1995 – July 1996 (part time).

Professor, Samuel Curtis Johnson Graduate School of Management, Cornell University (adjunct appointment), August 1987 – July 1995.



Chercheur étranger, Institut National de la Statistique et des Etudes Economiques (INSEE), Paris, Department of Research, August 1991 – July 1992, January 1993, January 1994.

Professeur visitant, HEC (Hautes Etudes Commerciales, Paris) Department of Finance and Economics, September 1991 – July 1992 and January 1993, December 1993 – January 1994.

Professeur visitant, CREST (Centre de Recherche en Statistique et Economie, Paris), September 1991 – July 1992, July 1993.

Associate Professor with tenure, Cornell University, August 1987 – December 1989.

Research Associate, Industrial Relations Section, Department of Economics, Princeton University, September 1986 – August 1987.

Visiting Associate Professor of Economics, Department of Economics, Massachusetts Institute of Technology, September 1985 – August 1986.

Associate Professor of Econometrics and Industrial Relations, Graduate School of Business, University of Chicago, September 1982 – August 1986. Assistant Professor, September 1979 – August 1982. Visiting Assistant Professor, September 1978 – August 1979.

Senior Study Director/Research Associate, NORC/Economics Research Center, 6030 Ellis Avenue, Chicago, Illinois 60637, September 1978 – August 1986.

Academic Consultant, Centre for Labour Economics, London School of Economics, January 1979 – April 1979.

Assistant Professor of Economics, Department of Economics, Princeton University, September 1977 – August 1979 (on leave September 1978 – August 1979). Lecturer in Economics, September 1976 – August 1977.

Associate Editor, *Journal of Business and Economic Statistics*, 1983 – 1989.

Editorial Board, *Journal of Applied Econometrics*, 1987 – 1989.

Associate Editor, *Journal of Econometrics*, 1987 – 1989.

## EDUCATION

Ph.D. Department of Economics, University of Chicago, December 1977.  
Thesis: An Econometric Model of the U.S. Market for Higher Education

M.A. Department of Economics, University of Chicago, March 1976.

A.B. Department of Economics (with highest honors), University of Notre Dame, May 1973.

## LANGUAGES

English (native), French

## HONORS AND FELLOWSHIPS

Fellow, American Association for the Advancement of Science (elected October 2020)

Julius Shiskin Award, American Statistical Association, Business and Economic Statistics Section (2016)

Cornell University, Graduate and Professional Student Assembly Award for Excellence in Teaching, Advising, and Mentoring (May 2015)

Fellow, Econometric Society (elected November 2014)

Roger Herriot Award, American Statistical Association, Government and Social Statistics Sections (2014)

Elected member, International Statistical Institute (March 2012)

Council of Sections (2014-2016), Chair (2013) Business and Economic Statistics Section (Chair-elect 2012), American Statistical Association

President (2014-2015), Society of Labor Economists, President-elect (2013-2014), Vice President (2011-2013)

Fellow, The American Statistical Association (elected August 2009)

Fellow, Society of Labor Economists (elected November 2006)

La bourse de haut niveau du Ministère de la Recherche et de la Technologie, fellowship for research at the Institut National de la Statistique et des Etudes Economiques (INSEE) awarded by the French Government, September 1991 – February 1992.

National Institute of mental Health postdoctoral fellow at NORC, September 1978 – August 1980.

National Institute of Mental Health pre-doctoral fellow at the University of Chicago, September 1973 – June 1976.

## TEACHING EXPERIENCE

### *Graduate:*

Microeconometrics using Linked Employer–Employee Data (CREST-ENSAE)  
Understanding Social and Economic Data (Cornell, co-instructor: Lars Vilhuber)  
Third-year Research Seminar I and II (Cornell)  
Seminar in Labor Economics I, II, and III (Cornell)  
Microéconomie des Données Appariées (CREST-GENES, in French)  
Microéconomie et Microéconomie du Travail (Université de Paris I, in French)  
Economie du Travail (Université de Paris II, in French)  
Economics of Compensation and Organization (Cornell)  
International Human Resource Management (Cornell)  
Corporate Finance (Hautes Etudes Commerciales, Paris)  
International Human Resource Management (HEC, Paris)  
Workshop in Labor Economics (Cornell)  
Economics of Collective Bargaining (Cornell)  
Executive Compensation (Cornell)  
Labor Economics (MIT)  
Labor and Public Policy (MIT)  
Applied Econometrics I, II (Chicago)  
Introduction to Industrial Relations (Chicago)  
Econometric Theory I (Chicago)  
Industrial Relations and International Business (Chicago)  
Workshop in Economics and Econometrics (Chicago)  
Econometric Analysis of Time Series (Princeton)  
Mathematics for Economists (Princeton)

### *Undergraduate:*

Understanding Social and Economic Data (Cornell, co-instructor: Lars Vilhuber)  
Introductory Microeconomics (Cornell)  
Economics of Employee Benefits (Cornell)  
Economics of Wages and Employment (Cornell)

Corporate Finance (Cornell)  
 Introduction to Econometrics (Princeton)  
 Microeconomics (Princeton)

## BIBLIOGRAPHY

### Books

1. Abowd, John M. and Francis Kramarz (eds.) *The Microeconometrics of Human Resource Management*, special issue of *Annales d'économie et de statistique* 41/42 (Paris: ADRES, January/June 1996).
2. Abowd, John M. and Richard B. Freeman (eds.) *Immigration, Trade and the Labor Market* (Chicago: University of Chicago Press for the National Bureau of Economic Research, 1991).

### Articles

1. McKinney, Kevin L., John M. Abowd, and John Sabelhaus, "United States Earnings Dynamics: Inequality, Mobility, and Volatility," In Raj Chetty, John N. Friedman, Janet C. Gornick, Barry Johnson, and Arthur Kennickel, eds., *Measuring the Distribution and Mobility of Income and Wealth*, (Chicago: University of Chicago Press for the National Bureau of Economic Research, 2021), forthcoming. [download preprint] [download chapter (open access)]
2. Abowd, John M. "Official Statistics at the Crossroads: Data Quality and Access in an Era of Heightened Privacy Risk," *The Survey Statistician*, Vol. 83 (January 2021):23-26. [download (open access)]
3. McKinney, Kevin L., Andrew S. Green, Lars Vilhuber and John M. Abowd "Total Error and Variability Measures for the Quarterly Workforce Indicators and LEHD Origin-Destination Employment Statistics in OnTheMap" *Journal of Survey Statistics and Methodology* (November 2020). [download arxiv preprint], DOI: <https://doi.org/10.1093/jssam/smaa029>, supplemental online materials DOI: <https://doi.org/10.5281/zenodo.3951670>
4. Abowd, John M., Ian M. Schmutte, William Sexton, and Lars Vilhuber "Why the Economics Profession Must Actively Participate in the Privacy Protection Debate," *American Economic Association: Papers and Proceedings*, Vol. 109 (May 2019): 397-402, DOI:10.1257/pandp.20191106. [download preprint]
5. Abowd, John M. and Ian M. Schmutte "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices," *American Economic Review*, Vol. 109, No. 1 (January 2019):171-202, DOI:10.1257/aer.20170627. [AER, ArXiv preprint, Replication information]
6. Weinberg, Daniel H., John M. Abowd, Robert F. Belli, Noel Cressie, David C. Folch, Scott H. Holan, Margaret C. Levenstein, Kristen M. Olson, Jerome P. Reiter, Matthew D. Shapiro, Jolene Smyth, Leen-Kiat Soh, Bruce D. Spencer, Seth E. Spielman, Lars Vilhuber, and Christopher K. Wikle "Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Secure the Future of the Federal Statistical System?" *Journal of Survey Statistics and Methodology* (2018) DOI:10.1093/jssam/smy023. [download, open access] [download preprint]
7. Abowd, John M., Ian M. Schmutte and Lars Vilhuber "Disclosure Limitation and Confidentiality Protection in Linked Data," in A.Y. Chun, M. Larson, J. Reiter, and G. Durrant (eds.) *Administrative Records for Survey Methodology* (New York: Wiley, forthcoming). [download preprint]
8. Abowd, John M., Kevin L. McKinney and Ian M. Schmutte "Modeling Endogenous Mobility in Earnings Determination," *Journal of Business and Economic Statistics* Vol. 37, Issue 3 (2019):405-418. DOI: 10.1080/07350015.2017.1356727. [download preprint] [JBES]
9. Abowd, John M., Francis Kramarz, Sébastien Perez-Duarte, and Ian Schmutte "Sorting between and within Industries: A Testable Model of Assortative Matching," *Annals of Economics and Statistics* 129 (March 2018): 1-32. NBER WP-20472. [download preprint] [programs] [data]
10. Abowd, John M., Kevin L. McKinney and Nellie Zhao "Earnings Inequality and Mobility Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data," *Journal of Labor Economics* 36, S1 (January 2018):S183-S300 DOI: 10.1086/694104. [download, not copyrighted] [download preprint]

11. Abowd, John M. "How Will Statistical Agencies Operate When All Data Are Private?" *Journal of Privacy and Confidentiality*, Vol. 7, Issue 3, Article 1 (2017). [download, open journal]
12. Haney, Samuel, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber "Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics," ACM SIGMOD 2017, DOI: 10.1145/3035918.3035940. [download]
13. Abowd, John M. and Kevin L. McKinney "Noise Infusion as a Confidentiality Protection Measure for Graph-based Statistics" *Statistical Journal of the International Association for Official Statistics* (2016) Vol. 32, No. 1, pp. 127-135, DOI: 10.3233/SJI-160958. [download article, open access] [download preprint]
14. Abowd, John M. and Ian Schmutte "Economic Analysis and Statistical Disclosure Limitation" *Brookings Panel on Economic Activity* (Spring 2015): 221-267. [download article and discussion, open access] [download preprint]
15. Schneider, Matthew J. and John M. Abowd "A New Method for Protecting Interrelated Time Series with Bayesian Prior Distributions and Synthetic Data," *Journal of the Royal Statistical Society, Series A* (2015) DOI:10.1111/rssa.12100. [download preprint]
16. Lagoze, Carl, William C. Block, Jeremy Williams, Lars Vilhuber, and John M. Abowd "Data Management of Confidential Data." *International Journal of Digital Curation* 8, no. 1 (2013): 265-278. doi:10.2218/ijdc.v8i1.259. [download preprint]
17. Abowd, John M. and Martha H. Stinson "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," *Review of Economics and Statistics*, Vol. 95, No. 5 (December 2013): 1451-1467. doi:10.1162/REST\_a\_00352. [download, not copyrighted]
18. Abowd, John M., Matthew J. Schneider and Lars Vilhuber "Differential Privacy Applications to Bayesian and Linear Mixed Model Estimation," *Journal of Privacy and Confidentiality*: Vol. 5: Iss. 1 (2013): Article 4. [download, open access]
19. Abowd, John M., Francis Kramarz, Paul Lengeremann, Kevin L. McKinney, and Sébastien Roux "Persistent Inter-Industry Wage Differences: Rent Sharing and Opportunity Costs," *IZA Journal of Labor Economics*, 2012, 1:7, doi:10.1186/2193-8997-1-7. [download, open access] [online Appendix]
20. Abowd, John M., Lars Vilhuber and William Block "A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs," in J. Domingo-Ferrer and I. Tinnirello, eds., *Privacy in Statistical Databases 2012*, LNCS 7556, pp. 216-225, (2012). [download, open access]
21. Abowd, John M. and Lars Vilhuber "Did the Housing Price Bubble Clobber Local Labor Markets When It Burst?" *American Economic Review Papers and Proceedings* Vol. 102, No. 3 (May 2012): 589-93, doi:pdfplus/10.1257/aer.102.3.589. [download preprint] [online Appendix] [data Readme] [data]
22. Abowd, John M., R. Kaj Gittings, Kevin L. McKinney, Bryce E. Stephens, Lars Vilhuber, and Simon Woodcock "Dynamically Consistent Noise Infusion and Partially Synthetic Data As Confidentiality Protection Measures for Related Time-series," Federal Committee on Statistical Methodology, Office of Management and Budget, 2012 Research Conference Papers. [download, open access, cited on May 21, 2012] [download archival copy].
23. Abowd, John M. and Matthew Schneider "An Application of Differentially Private Linear Mixed Modeling," ICDMW, pp. 614-619, 2011 IEEE 11th International Conference on Data Mining Workshops, 2011. [download, open access]
24. Kinney, Satkartar K. , Jerome P. Reiter, Arnold P. Rezek, Javier Miranda, Ron S. Jarmin, and John M. Abowd "Towards Unrestricted Public Use Business Micro-data: The Synthetic Longitudinal Business Database," *International Statistical Review*, Vol. 79, No. 2 (December 2011):362-84, doi:10.1111/j.1751-5823.2011.00153.x. [download, subscription required] [download preprint]
25. Abowd, John M. and Lars Vilhuber "National Estimates of Gross Employment and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail," *Journal of Econometrics*, Vol. 161 (March 2011): 82-99, doi: 10.1016/j.jeconom.2010.09.008. [download preprint] [data]
26. Abowd, John M., Bryce Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators" in T. Dunne, J.B. Jensen and M.J. Roberts, eds., *Producer Dynamics: New Evidence from Micro Data* (Chicago: University of Chicago Press for the National Bureau of Economic Research, 2009), pp. 149-230. [download, not copyrighted] [archival copy]

27. Abowd, John M., Kevin McKinney and Lars Vilhuber "The Link between Human Capital, Mass Layoffs, and Firm Deaths" in T. Dunne, J.B. Jensen and M.J. Roberts, eds., *Producer Dynamics: New Evidence from Micro Data* (Chicago: University of Chicago Press for the National Bureau of Economic Research, 2009), pp. 447-472. [download, not copyrighted] [archival copy]
28. Abowd, John M. and Lars Vilhuber "How Protective are Synthetic Data," in J. Domingo-Ferrer and Y. Saygun, eds., *Privacy in Statistical Databases*, (Berlin: Springer-Verlag, 2008), pp. 239-246. [download preprint]
29. Abowd, John M., Francis Kramarz and Simon Woodcock "Econometric Analyses of Linked Employer-Employee Data," in L. Mátyás and P. Sevestre, eds., *The Econometrics of Panel Data* (The Netherlands: Springer, 2008), pp. 727-760. [download preprint]
30. Abowd, John M., John Haltiwanger and Julia Lane "Wage Structure and Labor Mobility in the United States," in E. P. Lazear and K. L. Shaw, eds., *Wage Structure, Raises, and Mobility: International Comparisons of the Structure of Wages within and Across Firms* (Chicago: University of Chicago Press for the National Bureau of Economic Research, 2008), pp. 81-100. [download] [download preprint]
31. Machanavajjhala Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber "Privacy: Theory Meets Practice on the Map," International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436. [download preprint]
32. Abowd, John M. and Francis Kramarz "Human Capital and Worker Productivity: Direct Evidence from Linked Employer-Employee Data," *Annales d'Economie et de Statistique*, No. 79/80, (Juillet/Décembre 2005): 323-338. [download preprint]
33. Torra, V. J.M. Abowd and J. Domingo-Ferrer "Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment," in J. Domingo-Ferrer and Luisa Franconi (eds.) *Privacy in Statistical Databases* (Berlin: Springer-Verlag, 2006), pp. 233-242. [download preprint]
34. Abowd, John M., Francis Kramarz and Sébastien Roux "Wages, Mobility and Firm Performance: Advantages and Insights from Using Matched Worker-Firm Data," *Economic Journal*, Vol. 116, (June 2006): F245-F285. [download preprint]
35. Abowd, John M., Francis Kramarz and Sébastien Roux "Heterogeneity in Firms' Wages and Mobility Policies," in H. Bunzel, B.J. Christensen, G.R. Neumann and J-M. Robin, eds., *Structural Models of Wage and Employment Dynamics*, (Amsterdam: Elsevier Science, 2006), pp. 237-268. [download preprint]
36. Abowd, John M. and Lars Vilhuber "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers," *Journal of Business and Economics Statistics*, Vol. 23, No. 2 (April 2005): 133-152, *JBES* Joint Statistical Meetings invited paper with discussion and "Rejoinder" (April 2005): 162-165. [download preprint].
37. Abowd, John M., John Haltiwanger, Ron Jarmin, Julia Lane, Paul Lengermann, Kristin McCue, Kevin McKinney, and Kristin Sandusky "The Relation among Human Capital, Productivity and Market Value: Building Up from Micro Evidence," in *Measuring Capital in the New Economy*, C. Corrado, J. Haltiwanger, and D. Sichel (eds.), (Chicago: University of Chicago Press for the NBER, 2005), Chapter 5, pp. 153-198. [download, not copyrighted] [download preprint]
38. Abowd, John M. and Simon Woodcock "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data," in J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases* (Berlin: Springer-Verlag, 2004), pp. 290-297. [download preprint]
39. Abowd, John M., John Haltiwanger and Julia Lane "Integrated Longitudinal Employee-Employer Data for the United States," *American Economic Review Papers and Proceedings*, Vol. 94, No. 2 (May 2004): 224-229. [download preprint]
40. Abowd, John M. and Julia Lane "New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers," in J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases* (Berlin: Springer-Verlag, 2004), pp. 282-289. [download preprint]
41. Abowd, John M. and Francis Kramarz "The Costs of Hiring and Separations," *Labour Economics*, Vol. 10, Issue 5 (October 2003): 499-530. [download preprint]
42. Abowd, John M. "Unlocking the Information in Integrated Social Data," *New Zealand Economic Papers*, 0077-9954, Vol. 36, No. 1 (June 2002): 9-31. [download preprint]
43. Abowd, John M. and Orley Ashenfelter "Using Price Indices and Sale Rates to Assess Short Run Changes in the Market for Impressionist and Contemporary Paintings" in *The Economics of Art Auctions*, G. Mosetto and M. Vecco (eds.), (Milan: F. Angeli Press, 2002). [download preprint] [access book]

44. Abowd, John M. and Simon Woodcock "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), 215-277. [download preprint]
45. Abowd, John M., Bruno Crépon and Francis Kramarz "Moment Estimation with Attrition: An Application to Economic Models," *Journal of the American Statistical Association*, 96, No. 456 (December 2001): 1223-1231. [download preprint]
46. Abowd, John M., Francis Kramarz, David Margolis, and Kenneth Troske "The Relative Importance of Employer and Employee Effects on Compensation: A Comparison of France and the United States," *Journal of the Japanese and International Economies*. Vol. 15, No. 4, (December 2001): 419-436. [download preprint]
47. Abowd, John M. Julia Lane and Ronald Prevost "Design and Conceptual Issues in Realizing Analytical Enhancements through Data Linkages of Employer and Employee Data" in the *Proceedings of the Federal Committee on Statistical Methodology*, November 2000. [download preprint]
48. Abowd, John M., Francis Kramarz, David Margolis and Kenneth Troske "Politiques salariales et performances des entreprises : une comparaison France/Etats-Unis," *Economie et Statistique*, No. 332-333 (2000): 27-38. [Corporate Wage Policies and Performances: Comparing France with the United States] [download preprint]
49. Abowd, John M. and David Kaplan "Executive Compensation: Six Questions That Need Answering," *Journal of Economic Perspectives*, 13 (1999): 145-168. [Preprint and supplementary materials available at <http://hdl.handle.net/1813/56585>]
50. Abowd, John M., Patrick Corbel and Francis Kramarz "The Entry and Exit of Workers and the Growth of Employment: An Analysis of French Establishments" *Review of Economics and Statistics*, 81(2), (May 1999): 170-187. [download preprint]
51. Abowd, John M. and Francis Kramarz "Econometric Analysis of Linked Employer-Employee Data," *Labour Economics*, 6(March 1999): 53-74. [download preprint]
52. Abowd, John M., Hampton Finer and Francis Kramarz "Individual and Firm Heterogeneity in Compensation: An Analysis of Matched Longitudinal Employer-Employee Data for the State of Washington" in J. Haltiwanger et al. (eds.) *The Creation and Analysis of Employer-Employee Matched Data*, (Amsterdam: North Holland, 1999), pp. 3-24. [download preprint]
53. Abowd, John M. and Francis Kramarz "The Analysis of Labor Markets Using Matched Employer-Employee Data," in O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics*, Volume 3(B), Chapter 40 (Amsterdam: North Holland, 1999), pp. 2629-2710. [download preprint]
54. Abowd, John M. Francis Kramarz and David Margolis "High Wage Workers and High Wage Firms," *Econometrica*, 67(2) (March 1999): 251-333. [download preprint]
55. Abowd, John M. Francis Kramarz, Thomas Lemieux, and David Margolis "Minimum Wages and Youth Employment in France and the United States," in D. Blanchflower and R. Freeman (eds.) *Youth Employment and Joblessness in Advanced Countries* (Chicago: University of Chicago Press, 1999), pp. 427-472. [download] [download preprint]
56. Abowd, John M. and Francis Kramarz "Internal and External Labor Markets: An Analysis of Matched Longitudinal Employer-Employee Data" in J. Haltiwanger, M. Manser, and R. Topel (eds.) *Labor Statistics and Measurement Issues* (Chicago: University of Chicago Press, 1998), pp. 357-370. [download] [download preprint]
57. Abowd, John M., Francis Kramarz, David Margolis and Kenneth Troske "The Relative Importance of Employer and Employee Effects on Compensation: A Comparison of France and the United States," in *Comparaisons internationales de salaires* (Paris: Ministère du travail et des affaires sociales and INSEE, 1996), pp. 315-327.
58. Abowd, John M. and Laurence Allain "Compensation Structure and Product Market Competition," *Annales d'économie et de statistique*, (January/June 1996, No. 41/42): 207-217. [download preprint]
59. Abowd, John M., Francis Kramarz and Antoine Moreau "Product Quality and Worker Quality," *Annales d'économie et de statistique*, (January/June 1996, No. 41/42): 300-322. [download]
60. Abowd, John M. and Francis Kramarz "The Microeconometrics of Human Resource Management: International Studies of Firm Practices, Introduction and Overview," *Annales d'économie et de statistique*, (January/June 1996, No. 41/42): 1-9 (French), 11-19 (English).
61. Abowd, John M. and Francis Kramarz "Les Politiques Salariales : Individus et Entreprises" (Compensation Policies: Individuals and Firms), *Revue Economique* 47 (May 1996): 611-622.

[download preprint]

62. Abowd, John M. and Francis Kramarz "The Economic Analysis of Compensation Systems: Collective and Individual" in Norman Bowes and Alex Grey, eds. *Job Creation and Loss: Analysis, Policy and Data Development* (Paris: OECD, 1996), pp. 47-54.
63. Abowd, John M. and Michael Bognanno "International Differences in Executive and Managerial Compensation" in R.B. Freeman and L. Katz, eds. *Differences and Changes in Wage Structures* (Chicago: NBER, 1995), pp. 67-103. [download]
64. Abowd, John M. and Thomas Lemieux "The Effects of Product Market Competition on Collective Bargaining Agreements: The Case of Foreign Competition in Canada," *Quarterly Journal of Economics* 108 (November 1993): 983-1014.
65. Abowd, John M. and Francis Kramarz "A Test of Negotiation and Incentive Compensation Models Using Longitudinal French Enterprise Data," in J.C. van Ours, G.A. Pfann and G. Ridder, eds. *Labour Demand and Equilibrium Wage Formation Contributions to Economic Analysis* (Amsterdam: North-Holland, 1993), pp. 111-46. [download preprint]
66. Abowd, John M. and Richard B. Freeman "Introduction and Summary" in J.M. Abowd and R.B. Freeman, eds. *Immigration, Trade and the Labor Market* (Chicago: NBER, 1991), pp. 1-25. [download]
67. Abowd, John M. and Thomas Lemieux "The Effects of International Competition on Collective Bargaining Outcomes: A Comparison of the United States and Canada," in J.M. Abowd and R.B. Freeman, eds. *Immigration, Trade and the Labor Market* (Chicago: NBER, 1991), pp. 343-67. [download]
68. Abowd, John M. "The NBER Trade and Immigration Data Files," in J.M. Abowd and R.B. Freeman, eds. *Immigration, Trade and the Labor Market* (Chicago: NBER, 1991), pp. 407-21. [download]
69. Abowd, John M. "Does Performance-based Compensation Affect Corporate Performance?" *Industrial and Labor Relations Review* 43:3 (February 1990): 525-73S. Reprinted in *Do Compensation Policies Matter?* R.G. Ehrenberg, ed. (Ithaca, NY: ILR Press, 1990), pp. 52-73.
70. Abowd, John M., George Milkovich and John Hannon "The Effects of Human Resource Management Decisions on Shareholder Value," *Industrial and Labor Relations Review* 43:3 (February 1990): 203S-236S. Reprinted in *Do Compensation Policies Matter?* R.G. Ehrenberg, ed. (Ithaca, NY: ILR Press, 1990), pp. 203-236.
71. Abowd, John M. "The Effect of Wage Bargains on the Stock Market Value of the Firm," *American Economic Review* 79:4 (September 1989): 774-800. (working paper title: "Collective Bargaining and the Division of the Value of the Enterprise.")
72. Abowd, John M. and Joseph Tracy "Market Structure, Strike Activity, and Union Wage Settlements," *Industrial Relations* 57:2 (Spring 1989): 227-50.
73. Abowd, John M. and David Card "On the Covariance Structure of Earnings and Hours Changes," *Econometrica* 57:2 (March, 1989): 411-45.
74. Vroman, Wayne and John M. Abowd "Disaggregated Wage Developments," *Brookings Papers on Economic Activity* (1:1988): 313-46.
75. Abowd, John M. and David Card "Intertemporal Labor Supply and Long Term Employment Contracts," *American Economic Review* 77:1 (March 1987): 50-68.
76. Abowd, John M. "New Development in Longitudinal Data Collection for Labor Market Analysis: Collective Bargaining Data," *American Statistical Association 1985 Proceedings of the Business and Economic Statistics Section* (Washington, DC: ASA, 1985). (invited paper)
77. Abowd, John M. and Arnold Zellner "Estimating Gross Labor Force Flows," *Journal of Business and Economic Statistics* 3 (July 1985): 254-283.
78. Abowd, John M. and Arnold Zellner "Application of Adjustment Techniques to U.S. Gross Flow Data," *Gross Flows in Labor Force Statistics*, edited by Paul Flaim and Carma Hogue, Bureau of the Census/Bureau of Labor Statistics Conference Volume (Washington, DC: GPO, 1985).
79. Abowd, John M. and Mark Killingsworth "Employment, Wages, and Earnings of Hispanics in the Federal and Nonfederal Sectors: Methodological Issues and Their Empirical Consequences," in *Hispanics in the U.S. Economy*, edited by G. Borjas and M. Tienda (New York: Academic Press, 1985), pp. 77-125.
80. Abowd, John M. "Economic and Statistical Analysis of Discrimination in Job Assignment," *Industrial Relations Research Association Proceedings of the Thirty-Sixth Annual Meetings* (Madison, WI: IRRRA, 1984), pp. 34-47. (invited paper)
81. Abowd, John M. and Mark Killingsworth "Do Minority/White Unemployment Differences Really Exist," *Journal of Business and Economic Statistics* 2 (January 1984): 64-72.

82. Abowd, John M. and Arnold Zellner "Estimating Gross Labor Force Flows," *American Statistical Association 1983 Proceedings of the Business and Economic Statistics Section* (Washington, DC: ASA, 1983), pp. 162-67.
83. Abowd, John M. and Mark Killingsworth "Sex Discrimination, Atrophy and the Male-Female Wage Differential," *Industrial Relations* 22 (Fall 1983): 387-402.
84. Abowd, John M. and Henry S. Farber "Job Queues and the Union Status of Workers," *Industrial and Labor Relations Review* 35 (April 1982): 354-67. [download]
85. Abowd, John M. and Orley Ashenfelter "Anticipated Unemployment, Temporary Layoffs and Compensating Wage Differentials," in *Studies in Labor Markets*, edited by S. Rosen (Chicago: University of Chicago Press for the NBER, 1981), pp. 141-170. [download]
86. Abowd, John M. "An Econometric Model of Higher Education," in *Managing Higher Education: Economic Perspectives*, A Monograph of the Center for the Management of Public and Nonprofit Enterprises (Chicago: University of Chicago Press, 1981), pp. 1-56.
87. Mulvey, Charles and John M. Abowd "Estimating the Union/Nonunion Wage Differential: A Statistical Issue," *Economica*, 47 (February 1980): 73-79.
88. Abowd, John M. and T. James Trussell "Teenage Mothers, Labor Force Participation, and Wage Rates," *Canadian Studies in Population* (1980): 33-48.

### Monographs

1. Abowd, John M., Martha H. Stinson and Gary Benedetto *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project*, November 2006. [download archival copy and Excel tables at <http://hdl.handle.net/1813/43929>]
2. Abowd, John M. and Michael Bognanno "The Center for Advanced Human Resource Studies Managerial Compensation Database: User's Guide," March 1991.
3. Abowd, John M. and Michael Bognanno "The Center for Advanced Human Resource Studies Managerial Compensation Database: Technical Guide," March 1991.
4. Abowd, John M. *An Econometric Model of the U.S. Market for Higher Education* (New York: Garland Press, 1984).
5. Abowd, John M. and Mark Killingsworth "Employment, Wages, and Earnings of Hispanics in the federal and Nonfederal Sectors," in *Hispanics in the Labor Force: A Conference Report*, edited by G. Borjas and M. Tienda. Final Report to the National Employment Policy Commission (Washington, DC: GPO, 1982).
6. Abowd, John M. "Program Evaluation: New Panel Data Methods for Evaluating Training Effects," in *Program Evaluation Final Report to the U.S. Department of Labor* (Contract No. 23-17-80-01) (Washington, DC: NTIS, 1983).
7. Abowd, John M. and Mark Killingsworth "Employment, Wages, and Earnings of Hispanics in the federal and Nonfederal Sectors," in *Hispanics in the Labor Force: A Conference Report*, edited by G. Borjas and M. Tienda. Final Report to the National Employment Policy Commission (Washington, DC: GPO, 1982).
8. Abowd, John M. "Minority Unemployment, Compensating Differentials and the Effectiveness of the EEOC," in *Issues in Minority and Youth Unemployment final Report to the U.S. Department of Labor* (Contract No. 20-17-80-44) (Washington, DC: NTIS, 1982)
9. Abowd, John M. and Mark Killingsworth "Structural Models of the Effects of Minimum Wages on Employment by Age Groups," *Final Report of the Minimum Wage Study Commission*, Volume 5 (Washington, DC: GPO, 1981).
10. Abowd, John M. and Mark Killingsworth "An Analysis of Hispanic Employment, Earnings and Wages with Special Reference to Puerto Ricans," *Final Report to the U.S. Department of Labor* (Grant 21-36-78-61) (Washington, DC: NTIS, 1981).

### Miscellany

1. Abowd, John M., Ian M. Schmutte, William Sexton, and Lars Vilhuber, *Introductory Readings in Formal Privacy for Economists* (May 8, 2019, updated regularly). [read, download]
2. Abowd, John M., "The Census Bureau Tries to Be a Good Data Steward in the 21st Century" International Conference on Machine Learning (ICML) 2019 keynote address. [video, start at minute 18:00] [slides]
3. Garfinkel, Simson L., John M. Abowd, and Christian Martindale, "Understanding Database Reconstruction Attacks on Public Data," *ACMQueue*, Vol. 16, No. 5 (September/October 2018): 28-53. [download, not copyrighted]



4. Garfinkel, Simson L., John M. Abowd and Sarah Powazek "Issues Encountered Deploying Differential Privacy," *WPES'18 Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, Ontario, CA (October 2018): 133-137, DOI:10.1145/3267323.3268949. [ArXiv preprint]
5. Abowd, John M. "The U.S. Census Bureau Adopts Differential Privacy," *KDD '18 Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, UK (August 2018): 2867, DOI:10.1145/3219819.3226070. [download, subscription required], [archival copy] [video]
6. Abowd, John M., Lorenzo Alvisi, Cynthia Dwork, Sampath Kannan, Ashwin Machanavajjhala, and Jerome Reiter "Privacy-Preserving Data Analysis for Federal Statistical Agencies," *Computing Community Consortium White Papers* (January 2017). [CCC white paper archive; ArXiv preprint]
7. Abowd, John M. "Why Statistical Agencies Need to Take Privacy-loss Budgets Seriously, and What It Means When They Do," presented to the Federal Committee on Statistical Methodology, Policy Conference, December 7-8, 2016. [download]
8. Vilhuber, Lars, John M. Abowd and Jerome P. Reiter "Synthetic Establishment Microdata around the World," *Statistical Journal of the International Association for Official Statistics*, Vol. 32 (2016): 65-68. [download, open access] [download preprint]
9. Abowd, John M. "Synthetic Establishment Data: Origins and Introduction to Current Research," *Statistical Journal of the International Association for Official Statistics*, Vol. 30, No. 2 (Summer 2014): 113-115. [download, subscription required] [download preprint]
10. Benedetto, Gary, Martha H. Stinson and John M. Abowd "The Creation and Use of the SIPP Synthetic Beta," U.S. Census Bureau Technical Paper (April 2013). [download]
11. Abowd, John M. and Lars Vilhuber "Science, Confidentiality, and the Public Interest," *Chance*, Vol. 24, No. 3 (Fall 2011): 58-62. [download]
12. Abowd, John M. "OnTheMap: Block-level Job Estimates Based on Longitudinally Integrated Employer-Employee Micro-data," *Association of Public Data Users Newsletter* Vol. 33, No. 2 (March/April 2010): 10-19. [download]
13. Abowd, John M. Kobbi Nissim and Chris Skinner "First Issue Editorial" *Journal of Privacy and Confidentiality*, Vol. 1, No. 1 (2009): 1-6. [download]
14. Abowd, John M. "Comments on 'Regional difference-in-differences in France using the German annexation of Alsace-Moselle in 1870-1918' by Matthieu Chemin and Etienne Wasmer" *NBER International Seminar on Macroeconomics* (2008): 306-309. [download]
15. Abowd, John M. and Julia Lane "The Economics of Data Confidentiality," *ICP Bulletin*, Volume 4, No. 2 (August 2007):18-21. [download preprint]
16. Abowd, John M. "Rapporteur comments: International Symposium on Linked Employer-Employee Data, Econometric Issues" *Monthly Labor Review* 121:7 (July, 1998): 52-53.
17. Abowd, John M. "Discussion of 'How much do immigration and trade affect labor market outcomes' by Geroge J. Borjas, Richard B. Freeman and Lawrence F. Katz." *Brookings Papers in Economic Activity*(1997:1): 76-82.
18. Abowd, John M. "Discussion of Gross Worker and Job Flows in Europe by M. Burda and C. Wyplosz." *European Economic Review*(1994): 1316-1320.
19. Abowd, John M. "Discussion of 'The Quality Dimension in Army Retention' by Charles Brown." in A. Meltzer (ed.) *The Carnegie-Rochester Conference on Public Policy* 33 (1990).
20. Abowd, John M. "Immigration, Trade, and Labor Markets in Australia and Canada," in *Immigration, Trade, and the Labor Market*, edited by R.B. Freeman (Cambridge, Mass: NBER, 1988), pp. 29-34.
21. Abowd, John M. "Discussion of 'Public Sector Union Growth and Bargaining Laws: A Proportional Hazards Approach with Time-Varying Treatments' by c. Ichniowski." in *Public Sector Unionism*, edited by R. Freeman (Chicago: University of Chicago Press for the NBER, 1988).
22. Abowd, John M., Ross Stolzenberg and Roseann Giarusso "Abandoning the Myth of the Modern MBA Student," *Selections The Magazine of the Graduate Management Admission Council* (Autumn 1986): 9-21.
23. Abowd, John M., Brent Moulton and Arnold Zellner "The Bayesian Regression Analysis Package: BRAP User's Manual Version 2.0," H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago, 1985.
24. Abowd, John M. and Mark R. Killingsworth "The Minimum Wage Law Winners and Losers," *The Wall Street Journal* (August 1981).

### Working and Unpublished Papers

1. McKinney, Kevin L. and John M. Abowd, "Male Earnings Volatility in LEHD before, during, and after the Great Recession," (August 2020). [download preprint]
2. Abowd, John M., Gary L. Benedetto, Simson L. Garfinkel et al. "The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau," (August 2020). [download preprint]
3. Abowd, John M., Ian M. Schmutte, William Sexton, and Lars Vilhuber "Suboptimal Provision of Privacy and Statistical Accuracy When They are Public Goods," (June 2019). [download preprint]
4. Abowd, John M., Joelle Abramowitz, Margaret C. Levenstein, Kristin McCue, Dhiren Patki, Trivellore Raghunathan, Ann M. Rodgers, Matthew D. Shapiro, Nada Wasi, 2019. "Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data," Working Papers 19-08, Center for Economic Studies, U.S. Census Bureau, handle: RePEc:cen:wpaper:19-08. [download preprint]
5. McKinney, Kevin L. Andrew Green, Lars Vilhuber, and John M. Abowd "Total Error and Variability Measures with Integrated Disclosure Limitation for Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in On The Map" (December 2017). [download preprint]
6. Abowd, John M. and Ian Schmutte "Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods" (April 2017), [download preprint], published as Abowd, John M. and Ian M. Schmutte "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices," *American Economic Review*, Vol. 109, No. 1 (January 2019):171-202, DOI:10.1257/aer.20170627. [AER, ArXiv preprint, Replication information]
7. Abowd, John M. "Where Have All the (Good) Jobs Gone? (May 2014) Society of Labor Economists Presidential Address. [download preprint] [accompanying audio]
8. Abowd, John M., John Haltiwanger, Julia Lane, Kevin McKinney and Kristin Sandusky "Technology and Skill: An Analysis of Within and Between Firm Differences" (March 2007) NBER WP-13043. [download preprint]
9. Abowd, John M., Francis Kramarz, David N. Margolis, and Thomas Philippon "Minimum Wages and Employment in France and the United States" (February 2006). [archival download]
10. Abowd, John M., Paul Lengermann and Kevin L. McKinney "The Measurement of Human Capital in the U.S. Economy," (March 2003) [download Census, cited on September 1, 2015] [archival download]
11. Abowd, John M., Robert Creecy and Francis Kramarz "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," (March 2002). [download Census, cited on September 1, 2015] [archival download] [Fortran source] [Support files] [VirtualRDC archive]

### MAJOR GRANTS AND RESEARCH CONTRACTS

1. Associate Director for Research and Methodology and Chief Scientist U.S. Census Bureau, Intergovernmental Personnel Act (IPA) with Cornell University, June 1, 2016—March 27, 2020.
2. Research and Methodology Support Services, U.S. Census Bureau contract with Cornell University, June 1, 2015—May 31, 2016, \$268,897.
3. The Economics of Socially Efficient Privacy and Confidentiality Management for Statistical Agencies, Alfred P. Sloan Foundation awarded to Cornell University, April 1, 2015—March 31, 2019, \$535,970. (co-PIs Lars Vilhuber and Ian Schmutte)
4. RCN: Coordination of the NSF-Census Research Network, National Science Foundation SES 1237602 awarded to the National Institute of Statistical Sciences, July 15, 2012—June 30, 2017, transferred to Cornell University, September 2014, \$748,577. (PI Lars Vilhuber, other co-PIs Alan Karr, Jerome Reiter)
5. NCRN-MN: Cornell Census-NSF Research Node: Integrated Research Support, Training and Data Documentation, National Science Foundation Grant SES 1131848 awarded to Cornell University, October 1, 2011—September 30, 2016, \$2,999,614. (with William Block, Ping Li, and Lars Vilhuber)
6. A Census-Enhanced Health and Retirement Study: A Proposal to Create and Analyze an HRS Dataset Enhanced with Characteristics of Employers, Alfred P. Sloan Foundation grant awarded to the Institute for Social Research, University of Michigan with a subcontract to

- Cornell University, September 1, 2011—August 31, 2016, Cornell component \$349,608. (PI: Margaret Levenstein; other co-PIs: Matthew Shapiro, Kristin McCue and David Weir)
7. Synthetic Data User Testing and Dissemination, National Science Foundation Grant SES 1042181 awarded to Cornell University, September 15, 2010 to September 14, 2013, \$197,170. (Co-PI Lars Vilhuber)
  8. CDI-Type II: Collaborative Research: Integrating Statistical and Computational Approaches to Privacy, National Science Foundation Grant BCS 0941226 awarded to Cornell University, September 1, 2010—August 31, 2014, \$409,296. (Other PIs: Aleksandra B Slavkovic, Stephen E. Fienberg, Sofya Raskhodnikova, and Adam Smith)
  9. TC: Large: Collaborative Research: Practical Privacy: Metrics and Methods for Protecting Record-level and Relational Data, National Science Foundation Grant TC 1012593 awarded to Cornell University, July 15, 2010 to July 14, 2015, \$1,326,660. (Other PIs: Johannes Gehrke, Gerome Miklau, and Jerome Reiter)
  10. The Longitudinal Employer-Household Dynamics Program, U.S. Bureau of the Census, Interagency Personnel Act (IPA) with Cornell University, September 18, 1998 – September 17, 2000, \$260,000; renewed September 14, 2000—September 13, 2002, \$320,000; contract renewed as consultant September 14, 2002—September 13, 2003 (\$120,000); renewed as IPA September 15, 2003 – September 14, 2005 (\$384,590); renewed as IPA September 15, 2005—September 14, 2007 (\$425,215); new September 15, 2008—September 14, 2010 (497,897); renewed September 15, 2010—September 14, 2012 (532,893); continued as a contract with ACES-Research, LLC (September 17, 2012–September 16, 2013); re-established as IPA October 1, 2013—September 30, 2014 (\$231,757); re-established as IPA November 14, 2014 –May 31, 2015 (\$229,095).
  11. Social Science Gateway to TeraGrid, National Science Foundation Grant SES 0922005 awarded to Cornell University, July 1, 2009 to June 30, 2012, \$393,523. (Co-PI Lars Vilhuber) [Cornell Chronicle Article] [ILR News Release]
  12. Joint NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection, July 2009 Washington, DC, National Science Foundation Grant SES 0922494 awarded to Cornell University, July 1, 2009 to June 30, 2010, \$18,480. (Co-PIs Lars Vilhuber, Jerome Reiter, and Ron Jarmin)
  13. The Economics of Mass Layoffs: Displaced Workers, Displacing Firms, Causes and Consequences, National Science Foundation Grant SES-0820349 awarded to Cornell University, October 1, 2008 to September 30, 2010, \$245,950. (Co-PI Lars Vilhuber)
  14. LEHD Developmental and Confidentiality Research, Census Bureau Contract to Abt Associates with subcontract awarded to Cornell University, August 1, 2007 to September 30, 2008, \$358,270.
  15. CT-T: Collaborative Research: Preserving Utility While Ensuring Privacy for Linked Data, National Science Foundation Grant CNS-0627680 awarded to Cornell University, September 5, 2006 to August 31, 2009, \$488,950. (PI Johannes Gehrke)
  16. LEHD Confidentiality Research, Census Bureau Contract to Abt Associates with subcontract awarded to Cornell University, October 1, 2004 to September 30, 2005, \$230,155.
  17. ITR-(ECS+ASE)-(dmc+int): Info Tech Challenges for Secure Access to Confidential Social Science Data, National Science Foundation Grant SES-0427889 awarded to Cornell University, October 1, 2004 to September 30, 2007, \$2,938,000. (Co-PIs Matthew D. Shapiro, Ronald Jarmin, Stephen F. Roehrig, and Trivellore Raghunathan) [Cornell Chronicle article]
  18. EITM: Developing the Tools to Understand Human Performance: An Empirical Infrastructure to Foster Research Collaboration, National Science Foundation Grant SES-0339191 awarded to Cornell University, October 1, 2004 to September 30, 2007, \$337,455 (Co-PIs John Haltiwanger and Ron Jarmin)
  19. The New York Research Data Center, National Science Foundation Grant SES-0322902 awarded to the NBER, August 1, 2003 to July 31, 2004, \$300,000. (PI Neil G. Bennett, Other co-PIs Bart Hobijn, Erica L. Groshen, Robert E. Lipsey)
  20. Workshop on Confidentiality Research, National Science Foundation Grant SES-0328395 awarded to the Urban Institute, June 1, 2003 – May 31, 2004, \$43,602. (Co-PI Julia Lane)
  21. Firms, Workers and Workforce Quality: Implications for Earnings Inequality and Economic Growth, Alfred P. Sloan Foundation Grant 22319-000-00 awarded to the Urban Institute, January 2003—January 2006, \$1,400,000. (Co-PIs John Haltiwanger, Julia Lane, J. Bradford Jensen, Fredrick Knickerbocker, and Ronald Prevost)
  22. The Demand for Older Workers: Using Linked Employer-Employee Data for Aging Research, National Institute on Aging, R01-AG18854-01 to Cornell University, July 1, 2002 – April 30,

- 2007, \$1,753,637. (Co-PIs John Haltiwanger, Andrew Hildreth, and Julia Lane)
- 23. Workers and Firms in the Low-wage Labor Market: Interactions and Long Run Dynamics, Russell Sage Foundation, Rockefeller Foundation, and Department of Health and Human Services (ASPE) to the Urban Institute \$700,000, September 1, 2001 August 31, 2003. (Co-PIs John Haltiwanger, Harry Holzer, and Julia Lane)
- 24. From Workshop Floor to Workforce Clusters: A New View of the Firm, Alfred P. Sloan Foundation, 99-12-12 to the Urban Institute, March 1, 2000 – March 31, 2002, \$314,604. (Co-PIs John Haltiwanger and Julia Lane)
- 25. Dynamic Employer-Household Data and the Social Data Infrastructure, National Science Foundation, SES-9978093 to Cornell University, September 28, 1999 – September 27, 2005, \$4,084,634. (Co-PIs John Haltiwanger and Julia Lane)
- 26. The Longitudinal Employer-Household Dynamics Program, National Institute on Aging, interagency funding to the United States Census Bureau, September, 1999 – August, 2001, \$490,000. Renewed September 2001– August 2004, \$750,000 (Co-PIs John Haltiwanger and Julia Lane) [Cornell Chronicle article]
- 27. Individual and Firm Heterogeneity in Labor Markets: Studies of Matched Employee-Employer Data, National Science Foundation SBR 9618111 to the NBER, March 15, 1997 – February 28, 2002, \$243,361.
- 28. Creation of an Employer Identification Link File and Addition of Employer Information to the National Longitudinal Survey of Youth 1979 Cohort, Bureau of Labor Statistics (subcontracted by NORC, University of Chicago, Chicago, IL 60637), July 1, 1995 – December 31, 1997, \$82,946.
- 29. Employment and Compensation Policies: Studies of American and French Labor Markets Using Matched Employer-Employee Data, National Science Foundation SBR 9321053 to the NBER, July 1, 1994 – June 31, 1997, \$ 185,257. (Co-PIs David Margolis and Kenneth Troske)
- 30. Compensation System Design, Employment and Firm Performance: An Analysis of French Microdata and a Comparison to the United States, National Science Foundation, SBR 9111186 to Cornell University, July 1, 1991 – December 30, 1994, \$174,565.
- 31. The Effects of Collective Bargaining and Threats of Unionization on Firm Investment Policy, Return on Investment, and Stock Valuation, National Science Foundation, SES 8813847 to the NBER, July 1, 1988 – June 30, 1990, \$81,107.
- 32. Improving the Scientific Research Utility of Labor Force Gross Flow Data, National Science Foundation, SES 85-13700 to the NBER, April 15, 1986 – March 31, 1988, \$69,993.
- 33. Program Evaluation: New Panel Data Methods for Evaluating Training Effects, U.S. Department of Labor Contract 23-17-80-01 to NORC at the University of Chicago, 1983.
- 34. Minority Unemployment, Compensating Differentials and the Effectiveness of the EEOC, U.S. Department of Labor Contract 20-17-80-44 to NORC at the University of Chicago, 1982.
- 35. An Analysis of Hispanic Employment, Earnings and Wages with Special Reference to Puerto Ricans, U.S. Department of Labor Grant 21-36-78-61, 1981.

**PROFESSIONAL SERVICE, SURVEYS, AND DATA COLLECTION**

- 1. Canadian Research Data Centre Network Inaugural Board 2017-2019.
- 2. American Economic Association, Committee on Economic Statistics (AEAWeb) 2013-2018.
- 3. National Academy of Sciences, Committee on National Statistics (CNSTAT) 2010-2013; reappointed 2013-2016.
- 4. National Academy of Sciences, CNSTAT, Panel on Measuring and Collecting Pay Information from U.S. Employers by Gender, Race, and National Origin, (Chair) 2011-2012.
- 5. National Academy of Sciences, CNSTAT, Panel on Measuring Business Formation, Dynamics and Performance, 2004-2007.
- 6. National Academy of Sciences, CNSTAT, Panel on Data Access for Research Purposes, 2002-2005.
- 7. Executive Committee, Conference on Research in Income and Wealth 2002-.
- 8. Distinguished Senior Research Fellow, LEHD Program, U.S. Census Bureau 1998-2016.
- 9. Social Science and Humanities Research Council (Canada), Major Collaborative Research Initiatives review panel, 1997, 1998.
- 10. Technical Advisory Board for the National Longitudinal Surveys of the Bureau of Labor Statistics, 1988-1990, 1992-2001, Chair 1999-2001.
- 11. National Science Foundation, Economics Panel, 1990-91, 1992-93; KDI Panel 1999; Infrastructure Panel 2000; CDI Panel 2008; CDI Panel 2009.

12. Principal Investigator for The Center for Advanced Human Resource Studies Managerial Compensation Data Base. sponsored by the Cornell University Center for Advanced Human Resource Studies, 1989-1994.
13. Principal Investigator for A Longitudinal Data Base of Collective Bargaining Agreements. Sponsored by the Bureau of National Affairs and the University of Chicago Graduate School of Business, 1985.

## PROFESSIONAL ORGANIZATIONS

1. American Economic Association
2. American Statistical Association
3. Econometric Society
4. Society of Labor Economists
5. International Statistical Institute
6. International Association for Official Statistics
7. National Association for Business Economics
8. American Association of Wine Economists
9. American Association for Public Opinion Research
10. Association for Computing Machinery
11. American Association for the Advancement of Science

## PERSONAL INFORMATION

United States citizen

Personal email: john.abowd@gmail.com

---

Hosted by CampusPress

**APPENDIX B — 2010 RECONSTRUCTION-ABETTED RE-IDENTIFICATION SIMULATED ATTACK**

1. This appendix provides a high-level summary of the reconstruction-abetted re-identification attack simulation that the Census Bureau conducted on the released 2010 Census data. To assess the risk of a reconstruction-abetted re-identification attack, the Census Bureau conducted a series of statistical exercises to quantify the contemporaneous and future risk that individual responses could be disclosed. The Census Bureau has completed two simulated attacks that address the re-identification risk of a 100% microdata file (a file with detailed, individual-level records for every person enumerated in the census) reconstructed from the published Summary File 1 data. The 2010 Summary File 1, usually called SF1, includes the 2010 P.L. 94-171 Redistricting Data Summary File, the 2010 Advanced Group Quarters Data Summary File, and the bulk of the demographic and housing characteristics released from the 2010 Census in tabular format.<sup>1</sup> The fundamental structure of these simulations is as follows.

**SIMULATED RECONSTRUCTION ATTACK**

2. Database reconstruction is the process of statistically re-creating the individual-level records from which a set of published tabulations was originally calculated. That is, database reconstruction attempts to “reverse engineer” the confidential input data used in a statistical tabulation system.
3. The Census Bureau released over 150 billion statistics as part of the 2010 Census. The simulated reconstruction attack used as its input a small fraction of those statistics—approximately 6.2 billion statistics contained in the following published SF1 tables from the 2010 Census:

P001 (Total Population by Block)  
P006 (Total Races Tallied by Block)  
P007 (Hispanic or Latino Origin by Race by Block)

---

<sup>1</sup> See the technical documents in [Summary File 1 Dataset \(census.gov\)](https://www.census.gov/data/tables/2010/sf1.html).

- P009 (Hispanic or Latino, and Not Hispanic or Latino by Race by Block)
- P011 (Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over by Block)
- P012 (Sex by Age by Block)
- P012A-I (Sex by Age by Block, iterated by Race)
- P014 (Sex by Single-year-of-age for the Population under 20 Years by Block)
- PCT012A-N (Sex by Single-year-of-age by Tract, iterated by Race)

4. The reconstruction of the 2010 Census microdata for the sex, age, race, Hispanic/Latino ethnicity, and census block variables was carried out by constructing a system of equations consistent with the published tables listed above that, once solved, could then be converted into microdata. This system of equations was solved using commercial mixed-integer linear programming software (Gurobi).
5. Because the parameters of the 2010 Census swapping methodology included invariants on total population and voting age population at the block level, the reconstruction was able to exactly reconstruct all 308,745,538 million records with correct block location and voting age (18+). Then, leveraging the race (63 categories), Hispanic/Latino origin, sex, and age (in years) data from the specified tables, the simulated attack was able to further reconstruct those variables on the individual-level records.
6. To assess the accuracy of these reconstructed individual-level records, the team performed exact record linkage of the five variables in the reconstructed microdata to the same five variables in the Census Edited File (CEF, the confidential data) and Hundred-percent Detail File (HDF, the confidential swapped individual-level data before tabulation). The results are summarized in Table 1. The “left” file of the record linkage is in the first column. The “right” file is the reconstructed microdata from SF1.

Table 1 Agreement Rates between the Reconstructed Microdata and the 2010 Census Edited File and Hundred-percent Detail File					
	Record Counts		Agreement Rates		
Left file	In Left	In Reconstructed	Exact	Fuzzy Age	One error
CEF	308,745,538	308,745,538	46.48%	70.98%	78.31%
HDF	308,745,538	308,745,538	48.34%	73.33%	80.39%

DRB clearance number CBDRB-FY21-DSEP-003

7. The agreement rates shown in Table 1 include block (which was never wrong), sex, age (in years), race (63 OMB categories), and Hispanic ethnicity and are computed as a percentage of the total population. Exact agreement means all five variables agreed precisely bit for bit. Fuzzy-age agreement means that block, sex, race, and Hispanic ethnicity agreed exactly, but age agreed only +/- 1 year (e.g., age 25 on the CEF is in fuzzy-age agreement with ages 24, 25, and 26 on the reconstructed data). The one-error agreement rate allows one variable – sex, age (outside +/- one year), race or ethnicity to be wrong.
8. Most errors in the reconstructed file are that the age variable is off by +/- 2 years rather than +/- 1 year. This error is the balance of the width of the 5-year categories used in the block-level summaries. Hence, even though the disclosure avoidance requirement for the 2010 Census SF1 tabular summaries specified block-level aggregation to 5-year bins for those age 20 and over, the effective aggregation was far less.
9. Figure 1 shows the distribution of agreement rates by block size. Agreement rates are only substantially lower than the population averages shown in Table 1 for blocks with populations between 0 and 9 people, which is where the Census Bureau has said it concentrated the swaps.<sup>2</sup> However, uniqueness on sex, age, race, and ethnicity is not limited to small population blocks. *This is one of the principal failures of the 2010 tabular disclosure avoidance methodology – swapping provided protection for households deemed “at risk,” primarily those in blocks with small populations, whereas for the entire 2010 Census a full 57% of the persons are population uniques on the basis of block, sex,*

---

<sup>2</sup> McKenna, L. (2018), “Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing,” <https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf>, p. 8.



age (in years), race (OMB 63 categories), and ethnicity. Furthermore, 44% are population uniques on block, age and sex.<sup>3</sup>

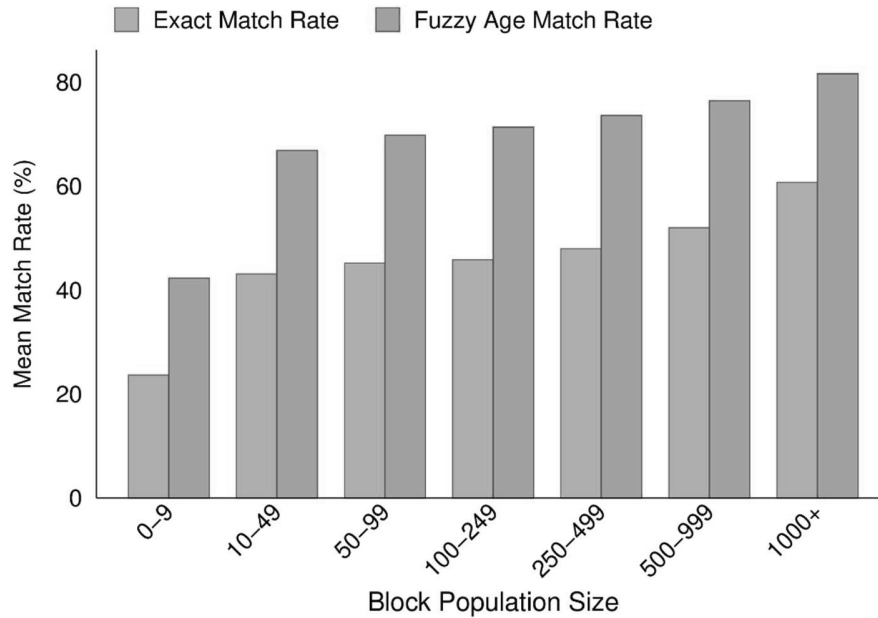


Figure 1 Block-level agreement rates between the reconstructed 2010 Census microdata and the 2010 Census Edited File by population in the block DRB clearance number CBDRB-FY21-DSEP-003.

10. Although there are no recent re-identification studies for decennial Public Use Microdata Samples (PUMS) with geography coded to the Public Use Microdata Area (PUMA), the Census Bureau continues to use 100,000 persons as the minimum population threshold for such areas and has coded geography on the 2010 PUMS and all American Community Survey (ACS) PUMS using these PUMAs. Since sex and age (single years) are population uniques at the tract level for only 0.18% of persons, this may still be justifiable for a 10% sample of 2010 Census records, but the potential re-

<sup>3</sup> The statistics in this paragraph are cleared for public release by the Census Bureau Disclosure Review Board (CBDRB-FY21-DSEP-003).

identification rate for a 100% public-use microdata file geocoded to the block level is certainly quite large.

11. The reconstruction experiment demonstrated that existing technology can convert the Census Bureau's traditional tabular summaries of Census data which was released in 2010 into a 100% coverage microdata file geocoded to the block level with very limited noise which was not released in 2010. This microdata file contains so much detail that it would have been deemed "unreleasable" if it had been proposed in conjunction with the original 2010 Census data products.
12. The ability to reconstruct the microdata means that there is now a significant disclosure risk for the 2010 Census Summary Files 1 and 2 (SF1, SF2) and the American Indian Alaska Native Summary File (AIANSF) data. There are approximately 150 billion statistics in the SF1, SF2, and AIANSF summaries (recall that the 2010 P.L. 94-171 Redistricting Data Summary File and the 2010 Advanced Groups Quarters Summary File are part of SF1). Because of the features noted above, releasing this many very accurate statistics made the ensemble of those publications equivalent to releasing the 2010 Hundred-percent Detail File (HDF), the swapped version of and the 2010 Census Edited File (CEF). There can be no uncertainty about this: *the 2010 Census tabular publications were equivalent to releasing every tabulation variable in the 2010 HDF in universe public-use microdata files without the hierarchical structure--person and household records can be fully reconstructed, but not directly linked to each other.* The team that demonstrated this vulnerability stopped after reconstructing person-level records for block, sex, age (in years), race (63 OMB categories), and Hispanic ethnicity because the vulnerability had been fully exposed mathematically and demonstrated empirically.
13. There are 308,745,538 (U.S. only) person records and 131,704,730 housing unit records in both the 2010 HDF and CEF, linked in their correct hierarchy. For the unswapped records in HDF, the images are identical to their CEF counterparts. For the swapped household records, the block identifier, household size, adult (age 18+) household

size, occupancy, and tenure variables are identical to their unswapped counterparts and on the person record the voting-age variable is identical to the unswapped counterpart.

14. As the documentation in McKenna (2018, 2019a) makes clear, a public-use microdata file containing the 308,745,538 person records in the HDF including only the five tabulation variables block, sex, age (in years), race (63 OMB categories), and Hispanic ethnicity is so disclosive that it would not have passed the disclosure avoidance criteria used for the 2010 Census Public-Use Microdata Sample.<sup>4</sup> Furthermore, the same file would not have passed the disclosure avoidance criteria applied to SF1 itself.<sup>5</sup> The official 2010 PUMS had a geographic population threshold of 100,000, collapsed categories to national population thresholds of 10,000, used partially synthetic data for the group quarters population, and “topcoding, bottom-coding, and noise infusion for large households.” The PUMS was sampled from the swapped version of the 2010 HDF, not the Census Edited File.

15. The additional disclosure avoidance methods used for the 2010 PUMS are explicitly noted on pages 2-1 and 2-2 of its technical documentation. The definition of a Public Use Microdata Area also explicitly references its confidentiality protection purpose:

“The Public Use Microdata Sample (PUMS) files contain geographic units known as Public Use Microdata Areas (PUMAs). To maintain the confidentiality of the PUMS data, a minimum population threshold of 100,000 is set for PUMAs. Each state is separately identified and may be comprised of one or more PUMAs. PUMAs do not cross state lines. (page 1-2, emphasis added)”

---

<sup>4</sup> McKenna, L. (2019a) “Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples,” Research and Methodology Technical Report available at [Disclosure Avoidance Techniques Used for the 1960 Through 2010 Census](#).

<sup>5</sup> McKenna, L. (2018)

16. This failure to apply microdata disclosure avoidance matters because the reconstructed 2010 microdata for block, sex, age (in years), race (63 OMB categories), and Hispanic ethnicity are a very accurate image of the HDF, and the HDF is a very accurate image of the CEF, which is the reason that it is also confidential. Consequently, the new technology-enabled possibility of accurately re-constructing HDF microdata from the published tabular summaries and the fact that those reconstructed data do not meet the disclosure avoidance standards established at the time for microdata products derived from the HDF demonstrate that the swapping methodology as implemented for the 2010 Census no longer meets the acceptable disclosure risk standards established when that swapping mechanism was selected for the 2010 Census.
17. Having demonstrated that a 100% microdata file can be successfully reconstructed from the published 2010 Census tabulations, the Census Bureau proceeded to use these reconstructed microdata to simulate a re-identification attack on those data.

#### **DE-IDENTIFICATION ATTACK SIMULATION**

18. The simulated re-identification attack proceeds as follows. Identify a person-level data source file that contains name, address, sex, and birthdate (e.g., commercially available data). Convert the names and addresses to their corresponding Census Bureau Protected Identification Key (PIK). Identify the corresponding census block for every address in the source file. Then, looping through all the records in the reconstructed microdata file produced from the reconstruction, find the first record in the source file that matches exactly on block, sex, and age. Once this step is completed, run through the remaining unmatched records from the reconstructed microdata and find the first unmatched record from the source file that matches exactly on block and sex, and matches on age plus or minus 1 year.
19. When both steps have been completed, output the records with successful matches from these two passes. These are called *putative re-identifications* because they appear

to link the reconstructed microdata to a real name and address associated with the block, sex, age, race, and ethnicity on the reconstructed microdata. These are the records the hypothetical attacker thinks are re-identified.

20. Putative re-identifications are not necessarily correct. An external attacker would have to do extra field work to estimate the *confirmation rate* – the percentage of putative re-identifications that are correct. An external attacker might estimate the confirmation rate by contacting a sample of the putative re-identifications to confirm the name and address. An external attacker might also perform more sophisticated verification using multiple source files to select the name and address most consistent with all source files and the reconstructed microdata.
21. At the Census Bureau we usually estimate the confirmation rate as a percentage of the total population, not as a percentage of the putative re-identifications, by performing a similar record linkage exercise of the putative re-identifications against the CEF, looking for exact matches on all variables (including PIK, block, sex, age, race, and ethnicity), followed by a second pass looking for exact matches except age, which is allowed to vary by plus or minus 1 year. Once these two passes have been completed, the matched records are the confirmed re-identifications, using exact match on PIK, block, sex, race (63 OMB categories), and ethnicity and match on age +/- 1 year as the definition of correct. The remaining unmatched records from the putative re-identifications of the reconstructed data are the unconfirmed re-identifications.
22. Table 2 shows the results of two such re-identification confirmation exercises. The first of these uses the combined commercial databases from Experian Marketing Solutions Incorporated, Infogroup Incorporated, Melissa Data Corporation, Targus Information Corporation, and VSGI LLC as the source file for name, address, sex, and age. This exercise simulates data quality circa 2010 for an external attacker relying on the consumer information in these databases. These results are in the row labeled “Commercial.” This re-identification experiment was the basis for the statistics released at the

American Association for the Advancement of Science 2019 annual meeting. Putative re-identifications were 138 million (45% of the 2010 Census resident population of the U.S.). Confirmed re-identifications were 52 million (17% of the same population).

23. Using the commercial data as the source for name, address, sex, and age is, as discussed in the main declaration, a best-case assumption. We know that these data exist and were available circa 2010 because that is when the Census Bureau acquired them. An external attacker, using the versions that the Census Bureau acquired and the relatively straightforward methodology above, would succeed at least as often as we did. This means that at least 52 million persons enumerated during the 2010 Census could be correctly re-identified using the attack strategy outlined here.
24. Suppose the external attacker had name, address, sex, and age of much better quality than the five commercial sources above. How much better could that attacker do using exactly the same strategy? This question can be answered by substituting the name, address, sex, and age from the 2010 CEF as the source file in the putative re-identification simulation. This is not cheating because no extra information in the CEF such as race, ethnicity or household structure is used for the source file. Hence, it is a proper worst-case scenario, and the one historically used by the Census Bureau in assessing microdata re-identification risk (see McKenna 2019b). If the external data on name, address, sex, and age are comparable to the 2010 Census, then the attacker will putatively re-identify 238 million persons (77% of the 2010 Census resident U.S. population). Confirmed re-identifications will be 179 million (58% of the same population). This means that with the best quality external data, relative to the 2010 Census, as many as 179 million persons could be correctly re-identified using the attack strategy outlined here.

Table 2 Record Linkage Summary from Commercial and CEF Record Sources				
PIK, Block, Age, Sex Record Linkage Source	Available Records	Records with PIK, Block, Sex, and Age	Putative Re-identifications using Source	Confirmed Re-identifications
Commercial	413,137,184	286,671,152	137,709,807	52,038,366
CEF	308,745,538	279,179,329	238,175,305	178,958,726
DRB clearance number CBDRB-FY21-DSEP-003.				

25. The record linkage results reported in Table 2 can be interpreted using two additional statistical quality measures: the *recall rate* and the *precision rate*. Taken together, these measures assess how successful an attacker can be at re-identifying records and how confident the attacker would be in those re-identifications.
26. *Recall rate*. The recall rate is the percentage of available source records that are correctly re-identified. Its numerator is the same as the confirmation rate, but its denominator is the number of records in the source file with sufficient information to perform the putative re-identification record linkage. For the two source files analyzed in these experiments, Table 2 shows the denominators for the recall rate in the column “Records with PIK, Block, Sex, and Age,” which gives the count of records with sufficient information to generate a putative match. Table 3 shows the recall rates for the two experiments. Both are greater than the respective confirmation rate because both the commercial data and the CEF have fewer usable records than the U.S. resident population. A critical result is the recall rate of 64% when the CEF is used as the source file. This result means that an external attacker with high quality name, address, sex, and age information succeeds in re-identification almost two times in three.

Table 3 Confirmation and Recall Rates		
Source	Percentage of U.S. Resident Population (Confirmation Rate)	Percentage of Complete Data Population (Recall Rate)
Commercial	16.85%	18.15%
CEF	57.96%	64.10%
DRB clearance number CBDRB-FY21-DSEP-003.		

27. *Precision rate.* Precision is the ratio of confirmed to putative re-identifications. It answers the question “How often is the attacker’s claimed re-identification correct as a percentage of the names the attacker attached to reconstructed census microdata?” Table 4 summarizes the precision rates for the two experiments. The precision of the experiment reported in February 2019 was 38% (first row of Table 4). The precision of the worst-case experiment is 75% (second row of Table 4). *This result means that an attacker using high-quality name, address, sex, and age data is correct three times out four.*

Table 4 Precision Rates	
Source	Confirmed Percentage of Putative Re-identification (Precision Rate)
Commercial	37.79%
CEF	75.14%
DRB Clearance number CBDRB-FY21-DSEP-003.	

28. To be successful, an attacker does not have to be a commercial entity, nor does a successful attack need to use commercially available data. Many agencies of federal, state and local governments in the U.S. now possess high-quality data on name, address,



sex, and age. When preparing public-use microdata files that contain variables that other agencies can access exactly, it has long been the practice to coarsen such data to prevent non-statistical uses by other agencies (see McKenna 2019b). Applying such precautions to decennial census data products would imply severe limitations on the variables published at the block level, even in the presence of swapping.

29. In conclusion, the Census Bureau's simulated reconstruction-abetted re-identification attack definitively established that the tabular summaries from the 2010 Census could be used to reconstruct individual record-level data containing the tabulation variables with their most granular definitions. Such microdata violated the disclosure avoidance rules that the Data Stewardship Executive Policy Committee had established for the 2010 Census and would not have been released had they been proposed as an official product because they posed too great a disclosure risk. The disclosure risk presumed by the 2010 standards recognized the excessive risk of re-identification if block geographic identifiers were placed on a 100% enumeration microdata file along with age (in years) and sex. The Census Bureau believed in 2010 that the minimum population that the geographic identifier could represent in such microdata is 100,000 persons – the size of a Public-Use Microdata Area. That belief was strongly confirmed by the simulated re-identification attack. Somewhere between 52 and 179 million person who responded to the 2010 Census can be correctly re-identified from the re-constructed microdata.

FINAL 2/16/10

**DSEP Meeting Record**

Topic: Updates

Meeting Date: 1/14/10

Attendees:

<i>Position</i>	<i>Attending for Position</i>
Deputy Director (Chair)	Tom Mesenbourg
AD, Administration	Ted Johnson
AD, Decennial	David Whitford
AD, Demographic	Cheryl Landman
AD, Economic	Harvey Monk
AD, Field	Marilia Matos
AD, IT	Brian McGrath
AD, Strategic Planning	Nancy Gordon
Rep. for Statistical Methodology	Tommy Wright
Senior Advisor for Data Management	Teresa Angueira
Chief, ITSO	
Chief, OAES	Kathleen Styles
Chief Privacy Officer	Mary Frazier
Also Attending:	Carol Comisarow, Ron Jarmin, David Raglin, Sharon Stern, Laura Zayatz, Michael Hawes

## **UPDATES**

### **Background**

[REDACTED]

### **Disclosure Review Board**

- The DRB has been using enhanced disclosure avoidance procedures and methods for projects involving sensitive topics and/or sensitive populations. These procedures were implemented in response to an August 2004 memo from Director Kincannon. Though the memo was superseded by the Custom Tabulations policy, the DRB was not informed of this, and has not changed its procedures for sensitive topics and populations.
- Laura Zayatz also voiced the DRB's concern about planning for the 2020 Census and continuing to release data at the block level, as block populations continue to decrease (e.g. 40% of blocks in North Dakota have only 1 household in them)

[REDACTED]

### **Action Items**

1. The DRB will develop recommendations for addressing the issue of disclosure review for sensitive populations. They will present their recommendations to DSEP once they have been vetted at the staff level.
2. DSEP agrees that the problem of block population size and disclosure avoidance is real, and that it deserves attention in the context of 2020 planning.

[REDACTED]

**DSEP Meeting Record**

Topics: [REDACTED]

Public Use File Reidentification Threats Update

Meeting Date: February 5, 2015

Attendees:

<i>Position</i>	<i>Attending for Position</i>
Deputy Director (Chair)	<i>absent</i>
AD, Administration	<i>absent</i>
AD, Decennial	Lisa Blumerman
AD, Demographic	Enrique Lamas
AD, Economic	<i>absent</i>
AD, Field	Jay Keller
AD, IT	Avi Bender
AD, Research and Methodology	Tom Louis
AD, 2020 Census	Lisa Blumerman
AD, Communications	Kim Collier
AD, Performance Improvement	Susan Reeves
Chief, PCO/ Chief Privacy Officer	Robin Bachman
Chief Demographer	Howard Hogan
Chief Information Security Officer	Tim Ruland
Asst. Director, Research and Methodology	Ron Jarmin
Also Attending:	Barbara Downs, Randy Becker, Byron Crenshaw, Laura McKenna, Heather Madray, Raj Dwivedy, Julie Atwell, Mike Castro

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

## **Public Use File Reidentification Threats Update**

### **Background and Discussion:**

On December 11, 2014, DSEP discussed a reidentification issue that occurred involving a Public Use File (PUF) produced as part of the New York City Housing Vacancy Survey (NYCHVS). At that meeting, DSEP commissioned a team to pursue the recommendations presented to DSEP in July 2014 in the paper titled “PUMS File Re-identification Threats and Potential Solutions for Mitigating those Threats.”

After discussing the logistics with some key stakeholders, and the difficulties of managing so many different angles on one team, DSEP approved a two-pronged approach to pursuing the paper’s recommendations.

The Center for Disclosure Avoidance Research (CDAR) has recently received authorization to hire new staff to focus primarily on synthetic data and reidentification research. This group is preparing a research proposal that focuses on the disclosure avoidance side of the PUF reidentification issue.

In addition to these efforts, the Demographic Programs Directorate (ADDP) will charter a team that focuses on the broader future of PUFs as well as some of the non-technical means of disclosure avoidance discussed in the July 2014 paper. This team will discuss Terms of Use for PUFs, restricted access, and other methods. This team will have representation from all of the impacted directorates. DSEP also recommended the team engage with external researchers on some of these ideas, and address their concerns.

### **Action Items:**

- CDAR will prepare a research proposal to outline future Census Bureau efforts in Synthetic Data and Reidentification Research.
- ADDP will charter a team to evaluate the future of PUFs and explore some of the non-technical solutions outlined in the July 2014 paper.

**DSEP Meeting Record**

Topics: Initial Request for DSEP Determination on Disclosure Avoidance for the 2018 End-to-End Test of the 2020 Census of Population and Housing (John Abowd, ADRM)

Record-level Re-identification Linkages for Evaluating the 2010 and 2020 Census Disclosure Avoidance Systems (John Abowd, ADRM)



Meeting Date: May 10, 2017

<i>Position</i>	<i>Attending for Position</i>
Deputy Director (Chair)	Laura Furgione
CAO	David Ziaya
CFO	Joanne Crane
AD, Decennial	Lisa Blumerman
AD, Demographic	Karen Battle
AD, Economic	Ron Jarmin
AD, Field	Joan Hill
AD, IT	Nitin Naik
AD, Research and Methodology	John Abowd
AD, 2020 Census	Lisa Blumerman
AD, Communications	Stephen Buckner
AD, Performance Improvement	Ted Johnson
Chief, PCO/ Chief Privacy Officer	Robin Bachman
Chief Demographer	Howard Hogan
Senior Advisor Designee from the Director's Office	<i>absent</i>
Chief Information Security Officer	<i>absent</i>

Asst. Director, Research and Methodology	John Eltinge
Also Attending:	Simson Garfinkel, Byron Crenshaw, Eloise Parker, Ashley Landreth, Mike Castro, Harold Saintelien, Janean Darden, Julie Atwell



## **Initial Request for DSEP Determination on Disclosure Avoidance for the 2018 End-to-End Test of the 2020 Census of Population and Housing**

### **Background:**

The Census Bureau's Research and Methodology Directorate (ADRM) is researching and developing disclosure avoidance methods and systems to replace those used for Census 2000 and the 2010 that were not designed to protect against database reconstruction attacks. ADRM is establishing the 2020 Disclosure Avoidance System (DAS), a formally private system based on the theoretical model known as differential privacy. This is the available technology for controlling reconstruction attacks.

The 2020 DAS team is working to establish adjustable formal privacy parameters for the 2018 End-to-End test. They are seeking DSEP concurrence with the Disclosure Review Board's (DRB's) April 10, 2017 determination that six data elements of PL 94-171 can continue to be published as enumerated. The team will test methods and systems with these elements published as enumerated for the 2018 End-to-End with the goal of making sound recommendations to DSEP for the full 2020 DAS. These elements to be published as enumerated are:

- the number of occupied housing units per block,
- the number of vacant housing units per block,
- the number of households per block,
- the number of adults (age 18+) per block (where the definition of an adult is inferred from the structure of the PL94-171 age categories),
- the number of children (age less than 18) per block (where the definition of a child is also inferred from the structure of the PL94-171 age categories),
- and the number of persons per block.

ADRM expects to perform follow-up analyses of the test products developed for the End-to-End Test. Because there is no national sample in 2018, some aspects of the differentially private system cannot be implemented in the End-to-End Test. They will have to be simulated from the 2010 Census data. This means that the demonstration data from the test can be made as noisy as DSEP wishes. However, there is only time to implement algorithms that maintain confidentiality with the six data elements in the 2010 PL94-171 redistricting data. There will be both policy and disclosure avoidance issues surrounding how broadly those products can be disseminated. Those issues will be brought to the DRB in a timely fashion.

ADRM also notes that DSEP will be asked to assume a formal policy consultant role for setting the confidentiality protection parameters for the final 2018 End-to-End Test and the 2020 DAS. The charter for DSEP currently delegates the authority to set disclosure avoidance standards to the DRB, with review by DSEP if necessary. However, these parameters now must be public in a formal privacy system. Furthermore, they, like any other operational decision need to be

discussed and set in a manner consistent with their importance in the publication of results from the 2020 Census. The privacy-loss setting recommended by DRB and DSEP, and accepted by the Director, will be implemented in the production system.

Requests to DSEP:

Request 1: Concurrence with the DRB's decision on the PL 94-171 file items that can be published as enumerated.

In order to meet the timeline for the 2018 End-to-End Test, the version of the DAS under development for the test is limited in scope to the PL94-171 redistricting data. ADRM will not have time to experiment with a suite of potential implementations. And, in particular, ADRM will not have time to modify certain implementation decisions. They will be put back on the table for the full 2020 Disclosure Avoidance System and the decision on these six specific items may be revisited.

Request 2: Concurrence with Change to DRB Operating Principles Related to 2020 Census

The second request is for DSEP concurrence on a change in the operating principles of the DRB for issues related to disclosure avoidance in the 2020 Census of Population. Because the differentially private disclosure avoidance methods operate on the ensemble of proposed publications, DSEP is asked to concur that any disclosure avoidance request for publications from 2020 Census data be routed to the 2020 DAS team first. Those requests should not be considered by the DRB until the 2020 DAS team supplies a memo stating that the requested publication can or cannot be incorporated into the total privacy-loss accounting.

This is not a request for a moratorium on approvals for decennial data releases or design. The privacy-loss budget itself and its allocation to various components of the publication system are policy decisions that the 2020 Disclosure Avoidance System team will not make. Those decisions will ultimately be made in a manner consistent with the charters of the DRB and DSEP, and defended by the Director.

There is very little historical guidance for this process. We need to develop practical use cases that illustrate the consequences of publication decisions under alternative privacy-loss scenarios. We need to document the extent to which a best-effort reconstruction of the 2010 Hundred-percent Detail File (HDF) is correlated with the actual HDF. This is going to take some time. In the interim, ADRM is asking the DRB to take a leadership role in making these important choices by enabling the development of technologies better adapted to global risk management.

Discussion:

DSEP recognized the value in ADRM's efforts to assemble a skilled team of experts in an effort to modernize Census Bureau disclosure avoidance techniques using formal privacy methods.

This is essential in light of research that demonstrates that we must protect against database reconstructions that could lead to re-identification.

DSEP discussed the details of the six data elements from PL 94-171 and considered the necessity of including all of these in the proposed 2020 DAS research. ADRM requested that all elements remain available for the 2018 test research with a reconsideration for the full 2020 DAS, once the Census Bureau understand the outcomes. Conversations with the Department of Justice for Voting Rights Acts requirements with PL 94-171 will also play a part in future decisions about published enumerations.

DSEP recognized the need to develop ways to communicate with state stakeholders and the public about data protections that based on 2020 DAS methods. Our messaging will have to provide some simpler description of how the methods make changes to the attributes of the people in block counts, but still provide accurate and usable data.

DSEP noted that The National Conference of State Legislators (NCSL) will be expecting updates from Decennial based on 2018 testing outcomes in anticipation of 2020 releases of PL 94-171. It will be important to engage NCSL in discussions about 2020 DAS methods.

DSEP acknowledged that this and other details from ADRM's research were scheduled for discussion at the May 10, 2017 meeting of the 2020 Census Portfolio Management Governing Board (PMGB). DSEP postponed further discussion on this project and requests, pending any feedback from the presentation on this topic to the 2020 PMGB.

#### **Post Meeting Notes:**

DSEP revisited this topic at the beginning of the May 11, 2017 meeting.

Regarding issues of surrounding Voting Rights Acts Requirements, DSEP recognized that Decennial would need to talk to Justice if we were to alter any of the 6 constraints from PL 94-171 for 2020.

DSEP noted that the 2020 PMGB is supportive of the efforts of the 2020 DAS to optimize output noise infusion methods while publishing the most accurate data possible. There was unanimous support from 2020 PMGB for DRB's determination that the six data elements from PL 94-171 should be published as enumerated and form the base for the 2018 End-to-End testing research with the 2020 DAS.

DSEP agreed that the DRB should require that any request for disclosure avoidance of proposed publications for the 2020 Census be routed to the 2020 DAS team before going to the DRB.

**Decision:**

Request 1: DSEP approves publication of the six data elements from PL 94-171 as enumerated for the 2018 End-to-End test. Based on lessons learned, the use of these constraints for the PL 94-171 will be revisited for 2020.

Request 2: DSEP agreed that the DRB should require that any request for disclosure avoidance of proposed publications for the 2020 Census be routed to the 2020 DAS team before going to the DRB.

**Record-level Re-identification Linkages for Evaluating the 2010 and 2020 Census Disclosure Avoidance Systems**

**Background:**

The DAS team is attempting a database reconstruction using data from the 2010 PL94-171 and SF1 tabulations. The next step is to link those reconstructed microdata to commercial name and address files obtained in support of post-2010 research meant to represent the type of publically available file an attacker might potentially acquire. These files include Experian, InfoGroup, Melissa, Targus, TransUnion, and VSGI. This linkage involves the use of name and address data.

The final step is to compare the fully reconstructed microdata, including the commercially supplied names and address, to the name and address data on the 2010 Census Unedited File (CUF). Following accepted disclosure avoidance evaluation practices on re-identification, the 2020 DAS team would report to DRB and DSEP the putative re-identification rate (percentage of the records in the reconstructed microdata that could be linked to name and address information in the commercial files) and the proportion of putative re-identifications that were correct (proportion of reconstructed data records with putative re-identifications that were correctly linked to 2010 Census responses, including name and address).

**Discussion:**

DSEP recognized that the project proposal meets Data Linkage Policy requirements and involves sensitive but critical work that will allow the 2020 DAS subteam to understand the degree of risk of re-identification and database reconstruction with Census files.

DSEP noted that the subteam assembled for this research is composed of federal employees and one SSS individual.

**Decision:**

DSEP approved this project.

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

**DSEP Meeting Record**

Topics: 2020 Decennial Record Linkage Test (Ned Porter, CSRM)



Meeting Date: May 11, 2017

<i>Position</i>	<i>Attending for Position</i>
Deputy Director (Chair)	Ron Jarmin
CAO	David Ziaya
CFO	Joanne Crane
AD, Decennial	Al Fontenot
AD, Demographic	Karen Battle
AD, Economic	Ron Jarmin
AD, Field	Joan Hill
AD, IT	Nitin Naik
AD, Research and Methodology	John Abowd
AD, 2020 Census	Al Fontenot
AD, Communications	Stephen Buckner
AD, Performance Improvement	Ted Johnson
Chief, PCO/ Chief Privacy Officer	Robin Bachman
Chief Demographer	Howard Hogan
Senior Advisor Designee from the Director's Office	<i>absent</i>
Chief Information Security Officer	Tim Ruland
Asst. Director, Research and Methodology	John Eltinge

Also Attending:	Simson Garfinkle, Tommy Wright, Eloise Parker, Ned Porter, Bill Winkler, Christa Jones, Letitia McKoy, Melissa Creech, Hampton Wilson, Ashley Landreth, Mike Castro, Janean Darden, Julie Atwell
-----------------	---

**Administrative Notes:**

At the beginning of the meeting, DSEP resumed their discussion and made a final decision on the topic: *Initial Request for DSEP Determination on Disclosure Avoidance for the 2018 End-to-End Test of the 2020 Census of Population and Housing*. The summary of that discussion and decision is in the May 10, 2017 meeting record.

**2020 Decennial Record Linkage Test**

**Background:**

Identifying duplicate records in the decennial census is critical to providing a more accurate count. One of the areas of research for improving the Decennial Matching methodology is improving the computer matching in the Duplicate Person Identification (DPI) process. This research will use the 2010 Census Unedited File (CUF) as well as data from Census Coverage Measurement (CCM). In addition, the research will determine if it is possible to increase the proportion of records receiving Personal Identification Keys (PIKs).

This research requires access to PIKs and complete name data on the files. This access will be limited to only five Census Bureau researchers as well as the Center for Statistical Research and Methodology's Data Steward. The data will be restricted to only authorized clusters.

**Discussion:**

DSEP acknowledged that research into deduplication methods is a routine and critical part of Census operations. DSEP further acknowledged that while this research project will use new technology and methods, it is fundamentally the same as research that happened in previous censuses.

**Decision:**

DSEP approved the project.

[REDACTED]



[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

**DSEP Meeting Record**

Topics:



Database Reconstruction Issue Mitigation (John Abowd, ADRM)

Meeting Date: February 15, 2018

<i>Position</i>	<i>Attending for Position</i>
COO (Chair)	Enrique Lamas
ADDC	Albert Fontenot
ADDP	Karen Battle
ADEP	Nick Orsini and Ron Jarmin
ADFO	Tim Olson
ADITCIO	Nitin Naik
ADRM	John Abowd
ADCOM	Stephen Buckner
CAO	David Ziaya
CFO	Joanne Crane
Asst. DRM	John Eltinge
Chief PCO/ Chief Privacy Officer	Robin Bachman
S.A. Director's Office	Douglas Clift
CISO	<i>Absent</i>
At-Large	Howard Hogan
At-Large	Frank Vitrano
Also Attending:	William Samples, David Waddington, Burton Reist, Victoria Velkoff, Robert Sienkiewicz, Jim Treat, Cynthia Hollingsworth, Clifford Jordan, Julia Naum, Jim Dinwiddie, Simson Garfinkel, Melissa Creech, Pat Cantwell, Byron Crenshaw, Hampton Wilson, Ashley Landreth, Mike Castro, Julie Atwell, Michael Snow

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

- [REDACTED]
- [REDACTED]

### **Database Reconstruction Issue Mitigation**

#### **Background**

The Census Bureau’s Operating Committee (OPCOM), serving as the Enterprise Risk Review Board, elevated the enterprise risk of database reconstruction to an enterprise issue based on the results of a database reconstruction attack research effort the Census Bureau launched to understand that risk better. When an enterprise risk is elevated to an enterprise issue, the risk owner must implement an active mitigation plan to mitigate the risk. To that end, the Research and Methodology Directorate presented six recommendations to help manage the Census Bureau’s publication strategy in ways that will protect its databases from reconstruction attacks.

NOTE: presenters and DSEP recognized that implementing several of the recommendations will require decisions on budget and staffing resources and that those decisions would need to be handled by other bodies at the Census Bureau. DSEP confined its discussion to establishing policy in response to the recommendations.

The following 6 recommendations were presented to DSEP:

- 1. Suspension until September 30, 2019 of ad hoc releases of sub-state geography from any confidential source unless vetted differential privacy tools, or a DRB-approved noise-infusion alternative, have been used to produce the publication. This applies to all research projects whether they are external or internal. It does not apply to scheduled publications from sponsored survey clients for whom there is already an approved DRB protocol. Those clients should be put on notice for subsequent contracts. The complete list of approved exceptions, including sponsored survey products, is provided in 20180215b-External\_Internal\_Substate\_Geography.xlsx. The suspension will be reviewed prior to September 30, 2019.**

NOTE: This suspension does not apply to state and national publications. It also does not apply to already scheduled publications from regular production activities. Program areas provided ADRM a list of those scheduled publications that should be exempted from the suspension. ADRM proposed ending those exemptions by September 30, 2019 even for those publications if they were not being produced using formally private systems by that point.

Discussion: DSEP recognized the need to modernize the Census Bureau's disclosure avoidance systems. DSEP acknowledged that by approving a list of exemptions they are agreeing to hold elevated levels of risk of database reconstruction associated with all of these data products. However, DSEP acknowledged the Census Bureau is obligated to provide the data the public needs for decision making and some of the release dates are required by law.

DSEP also acknowledged the need to set a target date for making these changes. While the ultimate goal is to make the publications of all of our programs formally private, that likely will not happen by September, 2019. However, in the meantime significantly improved noise infusion methods will be put in place to mitigate reconstruction risk.

DSEP members expressed concern that the list of already scheduled publications presented might be incomplete and asked for additional time for program areas to review the list and submit updates. DSEP agreed that the Center for Disclosure Avoidance Research (CDAR) should continue to accept submissions and finalize the list in advance of the next DSEP meeting. DSEP will formally approve the list at that point.

Decision: DSEP will finalize their approval of this recommendation at the March 15 DSEP meeting once the list of excepted publications has been finalized.

Action Items: Program areas will send updates on the table of exempted data releases to the Chief of CDAR by February 23. The Chief of CDAR will redistribute the combined list to all contributors by February 28. CDAR will finalize the list of approved exceptions for distribution before DSEP's meeting on March 15.

2. **Suspension of all proposed tables in Summary File 1 and Summary File 2 for the 2020 Census at the block, block-group, tract, and county level except for the PL94-171 tables, as announced in Federal Register Notice 170824806–7806–01 (November 8, 2017, pp. 51805-6). To add a summary file table at any level of geography, racial/ethnic subpopulation other than OMB aggregate categories as specified in the 1997 standard (Federal Register October 30, 1997, pp. 58782-90), or group quarters type below the 2010 P42 seven categories, an affirmative case must be made for that table, use cases identified, and suitability for use standards developed. In addition, we recommend that the voting-age invariant in PL94-171 be removed, so that voting-age would be protected. DSEP will be asked to approve the SF1 and SF2 table specifications once they have cleared 2020 governance.**

NOTE: The PL94-17 tables from the 2018 End-to-End Census Test have been designed with a formally private system already and will be published, with the voting-age invariant, as planned.

Discussion: DSEP recognized that the SF1 and SF2 involved a very detailed set of tables that had been created to suit a wide set of data users. These tables were created, as a rule, to produce as much highly accurate data as possible within the existing disclosure avoidance framework. However, DSEP acknowledged that these data in many cases were accurate to a level that was not supported by the actual uses of those data, and such an approach is simply untenable in a formally private system.

DSEP acknowledged a fundamental need to take stock of what data the Census Bureau is required to publish, both by statute and the needs of our data users, and at what level of accuracy. This is not an activity that should be done by our Disclosure Review Board. Program areas have to make the case of what the data will be used for, and the actual minimum level of accuracy needed for those uses, so that CDAR and the DRB can build the system to allocate the privacy-loss budget according to those use cases.

A redesign of SF1 and SF2 based on formally articulated use cases will take a tremendous amount of effort but cannot be done in a vacuum. Program areas will have to reach out to data-user communities on developing the use cases for the needed data accuracy and levels of geography.

NOTE: DSEP discussed but tabled until later any decision on changing the voting-age invariant for the PL94-171 table produced as part of the 2020 Census.



Decision: DSEP approved this recommendation. For the 2020 Census, SF1 and SF2 will be rebuilt based on use cases.

Action Items: DCMD, POP, and ADDC divisions will work with the relevant program management governing board (PMGB) to establish a plan to execute this redesign.

**3. Immediate review of all sub-state geography scheduled publications from the American Community Survey (ACS) to determine which ones can be delayed until there is a formally private publishing system for ACS.**

Discussion: DSEP acknowledged that many of the ACS tables are already in production and that production needs to move forward. DSEP acknowledged that there are likely no publications currently suitable for delay, however they emphasized that ACSO needs to ensure that all exceptions are added to the list.

Decision: DSEP approved this recommendation.

Action Items: ACSO will verify that they have included all of the necessary publications on the list of exempted data releases.

**4. Consideration of postponing ACS PUMS releases indefinitely.**

NOTE: DSEP recognized that all of the publication systems and methods for the Census of Island Areas are identical to the ACS. DSEP emphasized that any changes made to the ACS should also reflect consideration of the needs of the Island Areas.

Discussion: DSEP acknowledged that while the threat of database reconstruction and reidentification attacks applies to all of the Census Bureau's data products, should the ACS data be subject to a reidentification attack, from a public perception standpoint, our continued publication of the ACS PUMS files would appear to be an egregious mistake.

However, DSEP also acknowledged that the ACS PUMS is a heavily used dataset for research and recognized that discontinuing this publication could generate a great deal of traffic for the FSRDCs. DSEP acknowledged that, before the Census Bureau restricts use the ACS PUMS to the FSRDCs, it needs to verify that they can handle the increased workload. Additionally, at present there are no FSRDCs that are readily accessible from the Island Areas.

DSEP recognized that immediate suspension of the ACS PUMS would cause a great deal of concern among data users and others. DSEP discussed the need to work on messaging around

any suspension and to brief the Department of Commerce before the Census Bureau implements the suspension.

Decision: DSEP deferred for one month any decisions to suspend release of the ACS PUMS pending further consideration of the ability of the FSRDC network to support increased demand, the impact on the data needs of the Island Areas, and development of a messaging plan.

Action Items: ADRM will prepare an assessment of the potential increased demand on the FSRDC network, and Decennial will prepare an assessment of the impact of suspending this publication on the Island Areas. ADCOM will work on a messaging plan.

**5. Mandate for the 2022 Economic Censuses to use formally private publication systems for all tables.**

Discussion: DSEP recognized that it is too late to begin creating a formally private system for data releases from the 2017 Economic Census. DSEP additionally discussed how modernizing disclosure avoidance systems will involve much more than just budgeting extra funds. It also will require having the adequate number of people with the right skills to do the work.

DSEP recognized that program areas will have to involve their PMGB in setting resources, budgets, and timelines and that it should be feasible to put formally private systems in place in time for the 2022 Economic Census.

Decision: DSEP approved this recommendation. The Census Bureau will move forward with designing and implementing formally private systems for the 2022 Economic Census.

**6. Mandate to the Demographics Directorate to begin negotiations with survey clients for increased use of restricted-access microdata protocols and formally private table publication systems.**

POST MEETING NOTE: a member in attendance recommended that there should also be outreach to reimbursable clients for the Economic Directorate.

Discussion: DSEP recognized the need to begin discussions with sponsors of Census Bureau surveys but determined that the Census Bureau should have a communications plan in place before mandating that the Demographic Directorate speak to sponsors.

Decision: DSEP will reconsider in one month whether to mandate conversations with survey and report sponsors.

**Consolidated Action items:**

- Program areas will send updates on the table of exempted data releases to the Chief of CDAR by February 23.
- The Chief of CDAR will redistribute the combined list to all contributors by February 28.
- DCMD, POP, and the ADDC will work with the relevant PMGBs to establish a plan to execute the redesign of SF1 and SF2 based on use cases.
- ACSO will work to determine that all ACS data releases in production are listed on the spreadsheet of exceptions to the suspension.
- ADRM will prepare an assessment of the potential increased demand on the FSRDC network from suspension of the ACS PUMS.
- ADCOM will work on a messaging plan related to the suspension of the ACS PUMS.
- Decennial will prepare an assessment of the impact of suspending publication of the ACS PUMS on the Island Areas.

# Staring Down the Database

# Reconstruction Theorem

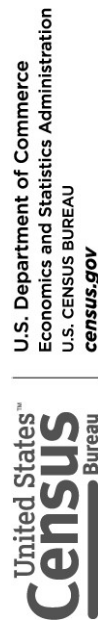
John M. Abowd

Chief Scientist and Associate Director for Research and Methodology

U.S. Census Bureau

American Association for the Advancement of Science

Annual Meeting Saturday, February 16, 2019 3:30-5:00

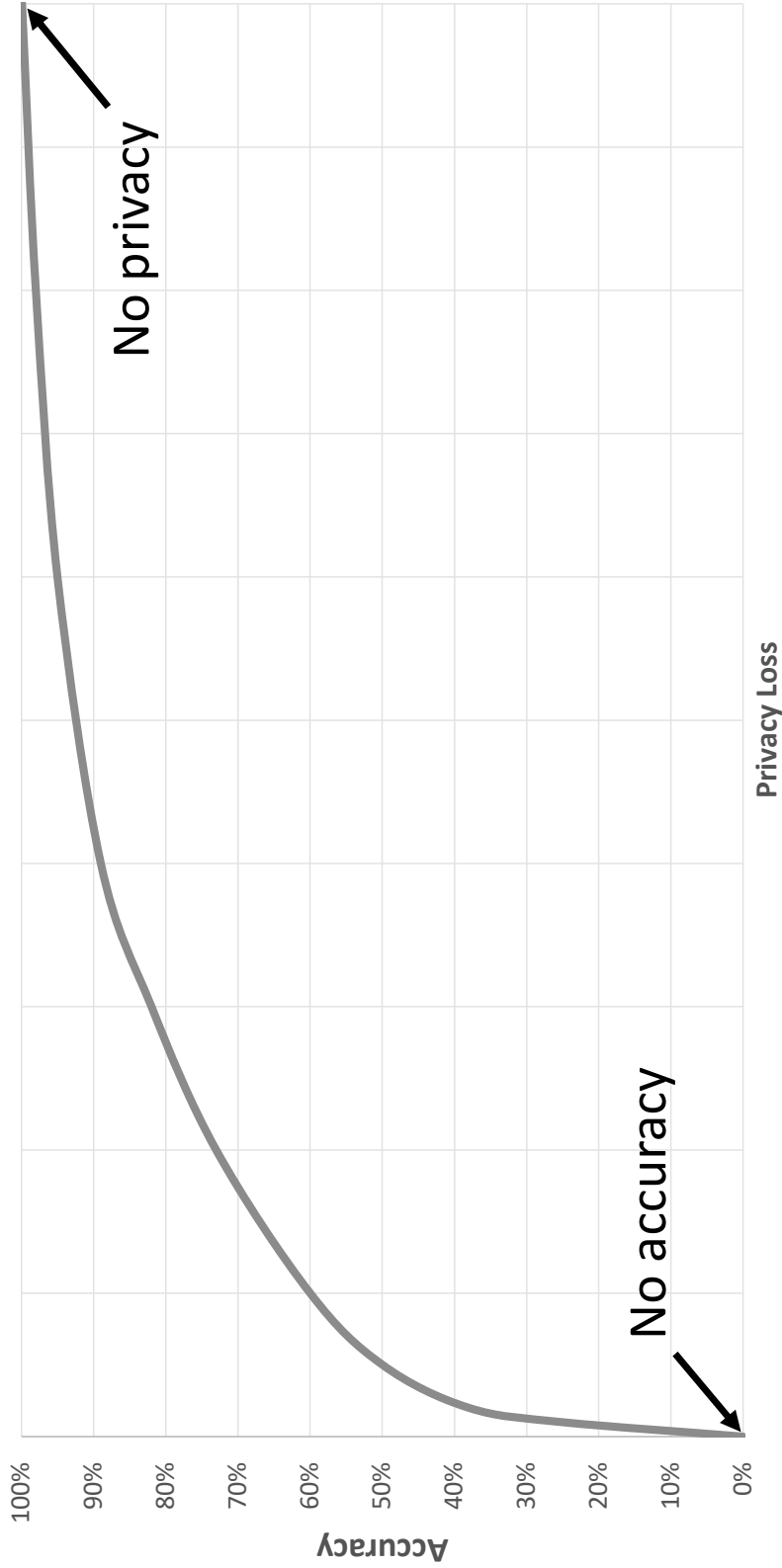


The views expressed in this talk are my own  
and not those of the U.S. Census Bureau.

- The challenges of a census:
1. collect all of the data necessary to underpin our democracy;
  2. protect the privacy of individual data to ensure trust and prevent abuse.

- Too many statistics
- Noise infusion is necessary
- Transparency about methods helps rather than harms

### Fundamental Tradeoff between Accuracy and Privacy Loss



# Good science and privacy protection are partners



# OnTheMap

Start Base Map Selection Results

## Distance/Direction Analysis Work to Home

### Display Settings

Labor Market Segment: **All Workers**  
 Filter: **All Workers**  
 Year: **2015**

### Map Controls

- Color Key
- Thermal Overlay
- Point Overlay
- Selection Outline
- Identify
- Clear Overlays
- Zoom to Selection
- Animate Overlays

### Report/Map Outputs

- Detailed Report
- Export Geography
- Print Chart/Map

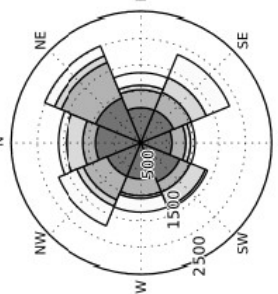
Legends

### Change Settings

Save Load Feedback Previous Extent Hide Tabs Hide Chart/Report



Job Counts by Distance/Direction in 2015  
All Workers



View as Radar Chart

## Jobs by Distance - Work Census Block to Home Census Block

	Count	Share
Total Primary Jobs	12,260	100.0%
Less than 10 miles	5,949	48.5%
10 to 24 miles	2,987	24.4%
25 to 50 miles	1,451	11.8%
Greater than 50 miles	1,873	15.3%

## What we did

- Database reconstruction for all 308,745,538 people in 2010 Census
- Link reconstructed records to commercial databases: acquire PII
- Successful linkage to commercial data: putative re-identification
- Compare putative re-identifications to confidential data
- Successful linkage to confidential data: confirmed re-identification
- Harm: attacker can learn self-response race and ethnicity

## What we found

- Census block correctly reconstructed in all 6,207,027 inhabited blocks
- Block, sex, age, race, ethnicity reconstructed
  - Exactly: 46% of population (142 million of 308,745,538)
  - Allowing age +/- one year: 71% of population (219 million of 308,745,538)
- Block, sex, age linked to commercial data to acquire PII
  - Putative re-identifications: 45% of population (138 million of 308,745,538)
- Name, block, sex, age, race, ethnicity compared to confidential data
  - Confirmed re-identifications: 38% of putative (52 million; 17% of population)
- For the confirmed re-identifications, race and ethnicity are learned exactly, not statistically

We fixed this for the 2020 Census by implementing differential privacy

## Acknowledgments

- The Census Bureau's 2020 Disclosure Avoidance System incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Scientist for Confidentiality and Data Access), Rob Sienkiewicz (ACC Disclosure Avoidance, Center for Enterprise Dissemination), Tamara Adams, Robert Ashmead, Michael Bentley, Stephen Clark, Craig Corl, Aref Dajani, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Edward Porter, Sarah Powazek, Anne Ross, Ian Schmutte, William Sexton, Lars Vilhuber, Cecil Washington, and Pavel Zhuralev

Thank you.

[John.Maron.Abowd@census.gov](mailto:John.Maron.Abowd@census.gov)

# More Background on the 2020 Census Disclosure Avoidance System

- September 14, 2017 CSAC (overall design)  
<https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf?#>
- August, 2018 KDD'18 (top-down v. block-by-block)  
<https://digitalcommons.ilr.cornell.edu/ldi/49/>
- October, 2018 WPES (implementation issues)  
<https://arxiv.org/abs/1809.02201>
- October, 2018 *ACMQueue* (understanding database reconstruction)  
<https://digitalcommons.ilr.cornell.edu/ldi/50/> or  
<https://queue.acm.org/detail.cfm?id=3295691>
- December 6, 2010 CSAC (detailed discussion of algorithms and choices)  
<https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#>

## Selected References

- Dinur, Irit and Kobbi Nissim. 2003. Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. In Halevi, S. & Rabin, T. (Eds.) Calibrating Noise to Sensitivity in Private Data Analysis Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg, 265-284, DOI: 10.1007/11681878\_14.
- Fellegi, Ivan P. 1972. On the Question of Statistical Confidentiality. Journal of the American Statistical Association, Vol. 67, No. 337 (March):7-18, stable URL <http://www.jstor.org/stable/2284695>.
- Ganda, Srivatsava, Shiva Kasiviswanathan and Adam Smith. 2008. Composition Attacks and Auxiliary Information in Data Privacy. In Knowledge, Discovery and Datamining, Las Vegas, NV, doi:10.1145/1401890.1401926.
- Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.
- McKenna, Laura. 2018. Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing, Working Papers 18-47, Center for Economic Studies, U.S. Census Bureau, Handle: RePEc:cen:wpaper:18-47.
- Ramachandran, Aditi, Lisa Singh, Edward Porter, and Frank Nagle. 2012. Exploring Re-Identification Risks in Public Domains, Tenth Annual International Conference on Privacy, Security and Trust, IEEE, doi:10.1109/PST.2012.6297917.
- U.S. Census Bureau. 2019. LEHD Origin-Destination Employment Statistics (2002-2015) [computer file]. Washington, DC: U.S. Census Bureau, Longitudinal-Employer Household Dynamics Program [distributor], accessed on February 15, 2019 at <https://onthemap.ces.census.gov>.



[Slide 1] [Before I start, I want to remind members of the audience that, while I am appearing in my official capacity as the Chief Scientist of the U.S. Census Bureau, I am presenting a summary of research findings. The views expressed in this talk are my own, not those of the Census Bureau.]

*Staring Down the Database Reconstruction Theorem*

[Slide 2] The 2020 Census will be the safest and best-protected ever. This is not nearly as easy as it sounds.

Throughout much of the history of the decennial census, our country has struggled with two challenges:

- 1) collect all of the data necessary to underpin our democracy;
- 2) protect the privacy of individual data to ensure trust and prevent abuse.

The first obligation derives directly from the Constitution, of course. As for the privacy requirement, Section 9 of the Census Act (Title 13 of the U.S. Code) prohibits making “any publication whereby the data furnished by any particular establishment or individual under this title can be identified.” In fact, the Census Bureau is about the only organization operating under a blanket U.S. legal requirement never to release data that can be tied back to individuals or companies no matter what.

The Census Bureau has always been committed to meeting both of its obligations; that is, providing population statistics needed by decision-makers, scholars, and businesses while also protecting the privacy of census participants.

A paper by Laura McKenna (2018), who supervised the confidentiality protection systems used by the Census Bureau for more than 15 years, catalogued the public information about the technical systems used for protection of publications from decennial censuses since 1970.

As McKenna noted, beginning with the 1990 Census, the primary confidentiality protection method employed was household-level swapping of geographic identifiers—moving an entire household from one location to another—prior to tabulating the data. The goal was to introduce uncertainty about whether households allegedly re-identified from the published data were correct.

Essentially the same methods were used for the 2000 and 2010 Censuses but with refinements that recognized the changing external environment.

The discipline of statistics has evolved over the last century. So too has the widespread availability of data. With each new development, the Census Bureau must ask how the current state of affairs will affect the production of the statistical products that it releases to the public so as to be both useful and privacy-preserving.

Sixteen years ago, two computer scientists, Irit Dinur and Kobbi Nissim (2003), wrote a seminal article proving a “database reconstruction theorem,” which is also known as the “fundamental law of information recovery.”

Three years later, Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith (2006) provided a mathematical foundation for what we now call “differential privacy.” In short, they explained how to quantify the limits on the accuracy of answers to queries based on the confidential data and the privacy-loss to the entities in those data, when the queries are answered publicly. More importantly, they provided a technique for enhancing privacy that goes far beyond the swapping approach that many statisticians have been using for years.

[Slide 3] The full implications of database reconstruction were not understood in 2003, but over the next several years a scientific consensus emerged in the data privacy community that:

- **Too many statistics**, published too accurately, expose the confidential database with near certainty (Dinur and Nissim 2003).
- **A necessary condition for controlling privacy loss** against informed attackers is to add noise to every statistic, calibrated to control the worst-case disclosure risk, which is now called a privacy-loss budget (Dwork, McSherry, Nissim and Smith 2006; Ganta, Kasiviswanathan, and Smith 2008).
- **Transparency about methods helps rather than harms**, Kerckhoff’s principle, applied to data privacy, says that the protections should be provable and secure even when every aspect of the algorithm and all of its parameters are public. Only the actual random number sequence must be kept secret (Dwork, McSherry, Nissim, and Smith 2006).

If you curate confidential data, then you can use those data for two competing goals:

- You can publicly and precisely answer statistical queries about the data.
- You can preserve and protect the privacy of those whose information is in the data.

You can do some of both.

[Slide 4] But if you do all of one, you can't do any of the other.

Period.

This trade-off is one of the hardest lessons to learn in modern information science. It is a lesson about data generally, not about counting people. And it is a mathematical theorem, not an opinion or implementation detail.

[Slide 5] This transformation in the fields of statistics and computer science is truly mind-blowing. It's at the heart of the science that we're here to celebrate. Cryptographers usually study the safety of methods for encrypting information about private data. Now their insights show us safe ways to publish information from private data. The cryptographic approach shows that some new methods can provably protect privacy, and some old methods provably do not. But the safe methods only work if we accept the inherent limitations on the accuracy of those publications that the cryptographers have highlighted.

Specifically, technical advances revealed a new vulnerability, allowing people to reconstruct data from tables that were previously assumed to be privacy-preserving, given the available computing resources. But other technical advances have also enabled a new form of privacy protection that is not only more sophisticated but also mathematically grounded in a way that allows statisticians to fully understand the limits of what they can make available and what kind of privacy they can provably offer. This dual breakthrough is transforming how we protect data today.

Good science and real privacy protection turn out to be partners, not competitors, in the efforts to modernize the methods data analysts use. For this reason, we have seen many companies, like Google, Microsoft, and Apple, turn to differential privacy to secure data and make guarantees about the privacy of

statistical tables. But it was actually the Census Bureau who first recognized the power of this method at scale.

[Slide 6] In 2008, the Census Bureau implemented an early version of differential privacy on data that display the commuting patterns of people based on where they live and work (Machanavajjhala et al. 2008; U.S. Census Bureau 2019).

Working with statisticians and computer scientists, we have collectively advanced the state of differential privacy such that we are going to implement it at scale as part of the 2020 Census. While I will talk about what that looks like in more detail tomorrow at 8:00AM, today I want to explain why we absolutely must implement differential privacy in order to protect the privacy of those participating in the census.

Starting in 1972, researchers began highlighting how it was possible to combine statistical tables and use differencing techniques to identify which census respondents provided the associated data (Fellegi 1972). As the market for detailed data grew and evolved, researchers also began highlighting how combining commercial data with census tables could introduce new vulnerabilities. While external users could not provably know whether or not their reconstructions were accurate, the Census Bureau recognized that it was critical to know the potential vulnerability of census data.

We acted proactively, as the Census Bureau has done for many decades. We designed our own internal research program to assess the current state of this vulnerability without waiting for a specific external threat. I'm now going to explain what we found.

[Slide 7] Here are the steps we followed:

- Using only published contingency tables (summary statistics), we applied the database reconstruction theorem to construct record-level images for all 308,745,538 persons enumerated in the 2010 Census. A record-level image is a row in the reconstructed database with the same variables that were used in publications from the confidential database. There is no traditional PII (personally identifiable information) on these reconstructed records.

- Using only the information in the reconstructed data records, we linked those records to commercial databases to acquire name and address information. This information would have been available to an external attacker, circa 2010.
- When the record linkage operation is successful, the PII from the commercial data are attached to the reconstructed census record. We call the reconstructed record, now laden with PII, “putatively re-identified,” which means that an attacker might think that the attack was successful.
- We then compared the putatively re-identified census records to the real confidential census records. When this comparison matched on all variables, including the PII and those variables not available in the commercial data, we called this a “confirmed re-identification.”
- The harm from such re-identifications, in the 2010 Census, is that the attacker learns the self-reported race and ethnicity on the confidential census record. Those data are not available in identifiable form to any commercial or governmental agency except the Census Bureau.

[Slide 8] Here are the basic results:

- In the reconstructed data, certain variables are always correctly reconstructed—meaning that the value in the reconstructed variable always matches its value in the confidential data. The census block, where the person lived on April 1, 2010, is always correctly reconstructed. This is true for every one of the 6,207,027 inhabited blocks in the 2010 Census.
- All the variables we studied: block, sex, age in years, race, and ethnicity are exactly correct in the reconstructed records for 46% of the population (142 million of 308,745,538 persons)—meaning that the reconstructed record exactly matches the confidential record on the value of all five variables. This result is salient because in the confidential data, more than 50% of the records are unique in the population—the only instance of this combination of values observed in the census (the exact percentage is confidential). If we allow the age to vary by plus or minus one year, then the number of reconstructed records that match the confidential data on these five variable rises to 71% (219 million of 308,745,538 persons).
- When we use the reconstructed block, sex and age to link each reconstructed record to the records harvested from commercial data

acquired at the time of the 2010 Census, we putatively re-identify 45% of the total population (138 million of 308,745,538 persons). That means that we were able to attach a unique name and address to 45% of the reconstructed records from the 2010 Census. The match is exact for block and sex. Age is allowed to vary by plus or minus one year.

- When we compared the unique name, block, sex, age, race, and ethnicity on the putative re-identifications to the same variables on the 2010 Census confidential data, we confirmed 38% of these matches (52 million of 308,745,538 persons, or 17% of the total population).

The putative re-identifications probably have a recall rate (or sensitivity) of at least 45%. Neither the attacker nor the Census Bureau have PII on all 308,745,538 persons enumerated in the 2010 Census, so the correct recall rate denominator is certainly less than the total population.

The precision of the record linkage is 38%, which means that the attacker would be correct between one-quarter and one-half of the time.

And both of these estimates (45% putatively re-identified; 38% of which are correct) are really lower bounds for other reasons: our experiments didn't use all of the information that the Census Bureau published from the 2010 Census. For example, we didn't use any information on household composition, which means that potential harm from discovering other features of households, like same-sex unions and adoptions, is still unquantified. We also made no use of the 2010 Public-Use Microdata Sample.

To further put these results in context, the last time the Census Bureau released results for a re-identification study, which did not use database reconstruction (Ramachandran et al. 2012), the putative re-identification rate was 0.017% (389 persons of 2,251,571) and the confirmation rate was 22% (87 of 389).

[Slide 9] All of us—the entire scientific community—have an obligation to examine the methods we use in light of the cryptographic critique of the privacy protections those methods offer. We must also recognize that these developments are sobering to everyone.

This is not just a challenge for statistical agencies or Internet giants, although those institutions have been in the vanguard of this movement.

It's a challenge for Internet commerce, because recommendation systems expose private data.

It's a challenge for bioinformatics, because summaries of genomes expose private data.

It's a challenge for commercial lenders, because benchmark risk assessments expose private data.

It's a challenge for nonprofit survey organizations, because their research reports expose private data.

Regardless of what anyone says, people want to be assured that their data are private. They want to know that we can't use statistical magic to re-identify information that they thought was private. They want to know that statistical tables can't come back to haunt them.

That's why I'm so grateful that the data we are showing today aren't the end of the story. They simply show that we cannot accept the status quo. We cannot presume that what worked a decade ago will work again in 2020. We have to innovate. And that's what we are doing.

In 2016, the Census Bureau acknowledged that database reconstruction was a vulnerability of the methods traditionally used to protect confidentiality in decennial census publications.

What we showed today is that we have a clear understanding of how it's possible to reconstruct 2010 Census data for block, sex, age, race and ethnicity. But this understanding isn't in vain. This understanding gave us the information we needed to develop techniques to make sure this isn't possible in 2020.

We are going into the 2020 Census confident that we can protect the privacy of all who participate. We have to make some important decisions about what statistics should be made available and how to weigh public data interests with our commitment to keep individual data private from reconstruction. But we know where the vulnerabilities are and we have the tools to make certain that what I showed today can't happen in the future.

The publications of the 2020 Census will be protected by differential privacy because it's imperative above all else that we ensure the trust of the American people.

The exact algorithms, and all parameters, will also be publicly released well in advance of the tables because it is imperative that we be accountable to the scientific community and the public at large.

[Slide 10] Statistics has evolved significantly over the last century. I'm honored to be a part of a statistical agency with a long tradition of implementing cutting-edge knowledge on the behalf of the American people. And I'm deeply grateful to the amazing team at the Census Bureau for identifying the challenges we face and ensuring that we can meet those challenges.

I promise the American people that they will have the privacy they deserve.

For those who would like to know more about how we are implementing differential privacy in the 2020 Census, please join me tomorrow at 8:00 AM where I will present our methods in more detail.

#### References

- Dinur, Irit and Kobbi Nissim. 2003. Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. In Halevi, S. & Rabin, T. (Eds.) Calibrating Noise to Sensitivity in Private Data Analysis Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg, 265-284, DOI: 10.1007/11681878\_14.
- Fellegi, Ivan P. 1972. On the Question of Statistical Confidentiality. Journal of the American Statistical Association, Vol. 67, No. 337 (March):7-18, stable URL <http://www.jstor.org/stable/2284695>.
- Ganda, Srivatsava, Shiva Kasiviswanathan and Adam Smith. 2008. Composition Attacks and Auxiliary Information in Data Privacy. In Knowledge, Discovery and Datamining, Las Vegas, NV, doi:10.1145/1401890.1401926.
- Machanavajhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.
- McKenna, Laura. 2018. Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing, Working Papers 18-47, Center for Economic Studies, U.S. Census Bureau, Handle: RePEc:cen:wpaper:18-47.



Ramachandran, Aditi, Lisa Singh, Edward Porter, and Frank Nagle. 2012. Exploring Re-Identification Risks in Public Domains, Tenth Annual International Conference on Privacy, Security and Trust, IEEE, doi:10.1109/PST.2012.6297917.

U.S. Census Bureau. 2019. LEHD Origin-Destination Employment Statistics (2002-2015) [computer file]. Washington, DC: U.S. Census Bureau, Longitudinal-Employer Household Dynamics Program [distributor], accessed on February 15, 2019 at <https://onthemap.ces.census.gov>.

UNITED STATES DISTRICT COURT  
FOR THE MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION

THE STATE OF ALABAMA, <i>et al.</i> ,	)	
	)	
Plaintiffs,	)	
	)	
v.	)	Civil Action No.
	)	3:21-CV-211-RAH
UNITED STATES DEPARTMENT OF	)	
COMMERCE, <i>et al.</i> ,	)	
	)	
Defendants.	)	

---

AMICUS BRIEF OF DATA PRIVACY EXPERTS

---

Ryan Calo  
Ran Canetti  
Aloni Cohen  
Cynthia Dwork  
Roxana Geambasu  
Somesh Jha  
Nitin Kohli  
Aleksandra Korolova  
Jing Lei  
Katrina Ligett

Deirdre K. Mulligan  
Omer Reingold  
Aaron Roth  
Guy N. Rothblum  
Aleksandra (Seša) Slavkovic  
Adam Smith  
Kunal Talwar  
Salil Vadhan  
Larry Wasserman  
Daniel J. Weitzner

Shannon L. Holliday  
(ASB-5440-Y77S)  
**COPELAND, FRANCO, SCREWS  
& GILL, P.A.**  
P.O. Box 347  
Montgomery, AL 36101-0347

Michael B. Jones  
Georgia Bar No. 721264  
jones@bmelaw.com  
**BONDURANT MIXSON &  
ELMORE, LLP**  
1201 West Peachtree Street, NW  
Suite 3900  
Atlanta, GA 30309

*Counsel for the Data Privacy Experts*

**TABLE OF CONTENTS**

**STATEMENT OF INTEREST..... 1**

**SUMMARY OF ARGUMENT ..... 1**

**ARGUMENT ..... 2**

**I. Reconstruction attacks Are Real and Put the Confidentiality of Individuals Whose Data are Reflected in Statistical Disclosures at Serious Risk ..... 2**

**A. Overview of Reconstruction Attacks..... 3**

**B. The Census Bureau’s Reconstruction Attack Demonstration ..... 4**

**C. Other Reconstruction Attack Demonstrations..... 7**

**D. Reconstruction Attacks Enable Re-Identification Attacks ..... 9**

**E. Reconstruction-Abetted Re-Identification Attacks Are a Realistic Threat ..... 11**

**II. Census Confidentiality Protections Must Evolve to Address Today’s Threats ..... 13**

**A. Differential Privacy is the Only Known Way to Protect Against Reconstruction Attacks ..... 13**

**B. The Census Bureau Cannot Tailor Its Confidentiality Protections to a Set of Predictable Risks as Suggested By Amicus Bambauer ..... 15**

**C. Heuristic Alternatives Have Several Limitations..... 16**

**III. Distinguishing the Census Bureau’s 2020 Disclosure Avoidance System (2020 DAS) and Differential Privacy ..... 16**

<b>IV. The 2020 DAS Does Not Use Statistical Inference .....</b>	<b>17</b>
<b>A. Differential Privacy As Used in the 2020 DAS Does Not Use Statistical Inference .....</b>	<b>18</b>
<b>B. The Post-Processing in Step Two of the DAS Does Not Use Statistical Inference .....</b>	<b>19</b>
<b>CONCLUSION .....</b>	<b>20</b>

### STATEMENT OF INTEREST

Amici are leading experts in data privacy and cryptography, and the connections of these fields to machine learning, statistics, and information theory. Amici's research and expertise with both differential privacy and the database reconstruction techniques used for reconstruction-abetted re-identification attacks offer a particularly well-informed perspective on the technical issues presented in this case.

### SUMMARY OF ARGUMENT

Amici, listed in Appendix A, submit this brief to provide the Court with a fuller understanding of the risks of reconstruction-abetted re-identification attacks, and the unique role that differential privacy plays in protecting statistical releases against them. This case is about the capacity of the Census Bureau to honor its confidentiality commitment in light of new and evolving threats. We offer the Court additional information about the prevalence of reconstruction attacks, the growing ease with which they can be undertaken, and the risks they pose to the privacy of census participants and therefore to the census and the important public purpose it serves.

Together, Amici have developed reconstruction attacks, proved that they are a mathematical certainty, and co-invented the only known methodology for addressing them. We write to assure this Court that the Census Bureau's decision to use differential privacy is sound and essential and reflects the widely shared understanding across the field that it is the only method available to protect statistical releases from reconstruction attacks. We also seek to clarify two technical points. *First*, differential privacy and the 2020 Disclosure Avoidance System (2020 DAS) are distinguishable. *Second*, based on available information about the 2020 DAS, it does not use statistical inference.

## ARGUMENT

### **I. Reconstruction Attacks Are Real and Put the Confidentiality of Individuals Whose Data Are Reflected in Statistical Disclosures at Serious Risk.**

To appreciate the critical need for the use of differential privacy in the protection of census data, it is vital for the Court to understand the threat posed by reconstruction attacks and the re-identification attacks they facilitate.

As the Census Bureau's research—as well as extensive academic research—shows, reconstruction and the more familiar re-identification attacks can go hand in hand: first, attackers reconstruct person-level data records from products based on aggregated personal data, then, re-identify the reconstructed records.<sup>1</sup> Data releases protected by traditional statistical disclosure limitation techniques are vulnerable to these attacks. Thus, traditional statistical disclosure limitation techniques are no longer adequate to meet the Census Bureau's obligation to maintain the confidentiality of individual census responses.

Although the data produced by reconstruction and re-identification attacks contain some misidentified records and uncertainty, the data still poses a threat. Many records resulting from such attacks are accurately identified. In addition, the risks of reidentification attacks are not evenly distributed throughout the population. Individuals with less common attributes or combinations of them are at greater risk of exposure.

---

<sup>1</sup> See *infra* Section I.

### A. Overview of Reconstruction Attacks.

Reconstruction attacks are processes for deducing highly accurate approximations of individual-level data from aggregated statistics.<sup>2</sup> There is a rich mathematical literature showing that reconstruction attacks pose a particular threat when the aggregated statistics consist of numerous simple counts like those published by the Census (e.g., the number of people in each census block broken down by race, ethnicity, and voting age).<sup>3</sup> The simple mathematical fact is that privacy loss from aggregate statistics works like radiation exposure: small, individually innocuous dosages of privacy erosion from published statistics accumulate until large-scale reconstruction is possible.

Reconstruction attacks reverse disclosure avoidance. Using only the information published in statistical reports, an attacker is able to deduce large swaths of the underlying confidential person-level data records with high accuracy.<sup>4</sup> These reconstructed records can then be identified—tied to an individual—by linking to commercial datasets using standard techniques. Therefore, if a purported disclosure avoidance technique allows reconstruction, then it does not meaningfully avoid disclosure.

Although reconstruction attacks are a new threat, carrying them out no longer requires significant innovation on the part of the attacker. The mathematical framework for reconstruction

---

<sup>2</sup> Cynthia Dwork, Adam Smith, Thomas Steinke & Jonathan Ullman, *Exposed! A Survey of Attacks on Private Data*, Annu. Rev. Stat. Appl. 4:12.1, 12.4 (2017).

<sup>3</sup> *Id.* at 12.4-12.6.

<sup>4</sup> *See generally id.* at 12.5-12.6.

attacks was discovered in 2003.<sup>5</sup> The technical community's understanding of such attacks was strengthened and generalized by subsequent work.<sup>6</sup>

### **B. The Census Bureau's Reconstruction Attack Demonstration.**

Aware of the developing literature on reconstruction attacks, the Census Bureau appropriately sought to gauge how "at risk" census data was to such attacks. As described in the declaration of Chief Scientist and Associate Director for Research and Methodology John M. Abowd, researchers from the Census Bureau revealed the results of a reconstruction attack on the 2010 Disclosure Avoidance System.<sup>7</sup> An internal research team was able to completely reconstruct much of the confidential data underlying statistical publications from the 2010 census.<sup>8</sup> Using only a small fraction of the summary statistical tables released to the public, the researchers were able to reconstruct the underlying database of "person-level" data with a high degree of accuracy, revealing records for all 308,745,538 individuals counted in the 2010 Census, including the

---

<sup>5</sup> Irit Dinur & Kobbi Nissim, *Revealing Information While Preserving Privacy*, Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 1, 202-10 (2003), <http://doi.acm.org/10.1145/773153.773173>.

<sup>6</sup> See generally Aloni Cohen & Kobbi Nissim, *Linear Program Reconstruction in Practice*, 10 J. of Privacy and Confidentiality 1 (2020), <https://doi.org/10.29012/jpc.711>; Cynthia Dwork, Frank McSherry, and Kunal Talwar, *The Price of Privacy and the Limits of LP Decoding*, STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, June 2007, at 85-94, [lpdecoding.pdf](http://kunal.org/lpdecoding.pdf) (kunal.org); Shiva Kasiviswanathan, Mark Rudelson, Adam Smith & Jonathan Ullman, *The Price of Privately Releasing Contingency Tables and the Spectra of Random Matrices with Correlated Rows*, STOC '10: Proceedings of the Forty-Second ACM Symposium on Theory of Computing, June 2010, at 775-784, <https://doi.org/10.1145/1806689.1806795>; Cynthia Dwork and Sergey Yekhanin, *New Efficient Attacks on Statistical Disclosure Control Mechanisms*, Advances in Cryptology – CRYPTO 2008, 1, 469-80 (David Wagner ed., 2008), [https://doi.org/10.1007/978-3-540-85174-5\\_26](https://doi.org/10.1007/978-3-540-85174-5_26).

<sup>7</sup> Decl. of John M. Abowd ¶ 38, Doc. 41-1.

<sup>8</sup> *Id.*



individual's "[b]lock, sex, age, race, [and] ethnicity."<sup>9</sup> The attack accurately reconstructed these fields for approximately 219 million people, or 71% of the population (to within one year of age), with exact reconstruction on 46% of the population, or 142 million people.<sup>10</sup> The Census Bureau's reconstruction demonstration closely followed the framework described by Dinur and Nissim. As described *infra*, the Census Bureau used the reconstructed data and commercially available datasets to successfully exactly reconstruct and re-identify the confidential census responses of 52 million people, without using any confidential Census data.

Using publicly and commercially available data and using eighteen-year-old techniques, the 2010 decennial census responses of 52 million people were reconstructed and re-identified. This is more than the combined 2010 enumerated population of the States of Alabama, Texas, and Florida—the Plaintiff's and the two largest amici states, respectively. If they are as vulnerable as the average census respondent, reconstruction and re-identification would expose the private Census responses of 91 members of the United States Congress, 23 members of the Alabama Legislature, and 24 sitting federal district and appellate judges within the Eleventh Circuit.

Plaintiff, Amicus Curae Jane Bambauer, and expert witness Steve Ruggles downplay the seriousness of this demonstration. The latter contrasts the Census Bureau's reconstruction results with a "simple simulation" of what can be predicted through chance alone.<sup>11</sup> Ruggles asks: What fraction of the population's 2010 census responses could be randomly guessed, rather than reconstructed? But Ruggles' analysis compares only on coarse statistics and vastly understates the real effectiveness of the Census attack.

---

<sup>9</sup> *Id.* at ¶ 38 & 108.

<sup>10</sup> *Id.*

<sup>11</sup> Ruggles' Expert Report, Appendix A, Page 7.

Ruggles does not separate out the rate of reconstruction according to Census block size.<sup>12</sup> The Census Bureau reconstruction does surprisingly well even on blocks of size 0-9: 20+% success; it achieves over 40% exact matches on blocks of size 10-49.<sup>13</sup> More than 32% of Alabama residents live in blocks of size 10-49.<sup>14</sup> On these blocks, Ruggles' random guessing has an inferior success rate of 12-15%. On blocks of size 0-9 its success rate is 3.5%. These comparisons are generous to Ruggles' random guessing; for example, census reconstructs age, sex, race, ethnicity, and block. Ruggles' guessing algorithm is given the block and guesses only age and sex.

Differential Privacy says that the risk of any harm remains essentially unchanged, independent of whether one joins or refrains from joining a dataset."<sup>15</sup> For residents of large blocks, participation in the Census will not substantially affect their likelihood of being reconstructed via random guessing. But without Differential Privacy, residents of small blocks will indeed suffer increased risk of reconstruction by participating in the Census. The Census Bureau has an obligation to protect the more than 1.7 million Alabamans living in small blocks.

---

<sup>12</sup> The size of a block matters. It's easier to randomly guess a card in your opponent's hand when playing Thirteen Card Rummy than when playing Three Card Poker. In the same way, random guessing works very well in blocks with hundreds or thousands of people, but very poorly for blocks with just tens of people.

<sup>13</sup> Appendix B of Declaration of John Abowd, Figure 1.

<sup>14</sup> *Id.*

<sup>15</sup> "The Mete and Measure of Privacy", Lecture by Cynthia Dwork, 152<sup>nd</sup> Annual Meeting of the National Academy of Sciences, April 2015, Research Briefings: April 25, 2015 (nasonline.org). The mathematical consequence is that algorithms operating on datasets should behave similarly on datasets that differ in the data of a single individual. How the *algorithm behaves* has nothing to do with what a privacy *adversary knows*, so Differential Privacy automatically protects against arbitrarily knowledgeable adversaries. This is the worst-case protection that Bambauer derides.

These small blocks are exactly those where attacks are most problematic, and where the protections of differentially private methods are most meaningful.

Reconstruction (and subsequent re-identification) of Census records does not require access to confidential Census records nor the expertise and computational resources of a federal agency. Columbia University Professor of Journalism Mark Hansen, working with a graduate student in statistics, “were able to perform our own reconstruction experiment on Manhattan. Roughly 1.6 million people are divided among 3,950 census blocks—which typically correspond to actual city blocks. The summary tables we needed came from the census website; we used simple tools like R and the Gurobi Optimizer; and within a week we had our first results.”<sup>16</sup> This attack used an academic version of Gurobi;<sup>17</sup> a commercial version would be much faster.

### **C. Other Reconstruction Attack Demonstrations.**

The same approach was used in 2018 to power another reconstruction attack, this one against a commercial disclosure avoidance system called Diffix.<sup>18</sup> Diffix was advertised as a system that provides off-the-shelf compliance with Europe's General Data Protection Regulation (GDPR).<sup>19</sup> According to its creators, “the French national data protection authority” had evaluated Diffix and found that it “delivers GDPR-level anonymity.”<sup>20</sup> Cohen and Nissim adapted the 2003 reconstruction attack blueprint to Diffix and, with a few hundred lines of code running in less than

---

<sup>16</sup>Mark Hansen, *To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data*, N.Y. Times (Dec. 5, 2018), <https://www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html>.

<sup>17</sup> *Id.*

<sup>18</sup> Cohen & Nissim, *supra* note 7 at 3-4.

<sup>19</sup> *Id.*

<sup>20</sup> *Id.*

ten seconds on a laptop, perfectly reconstructed the data without any error.<sup>21</sup> What happened next is typical of the patch-break-patch again cycle that plagues traditional approaches to disclosure limitation that do not have rigorous guarantees: the company behind Diffix updated their system and claimed to defend against the attack.<sup>22</sup> However, it was quickly shown that a slight modification of the same attack could still perfectly reconstruct the social security numbers of about 90% of data subjects.<sup>23</sup>

A reconstruction attack on statistical reports released by the Israel Central Bureau of Statistics (CBS) was carried out in 2014.<sup>24</sup> CBS conducts an annual Social Survey, with questions about religion, ethnicity, employment, education, income, family, health, and attitudes, among many others.<sup>25</sup> CBS made these statistical reports publicly available online.<sup>26</sup> Undergraduate computer science students demonstrated that they could completely reconstruct the survey responses of over 14% of data subjects—1005 out of the 7064 survey subjects.<sup>27</sup> The students stopped reconstructing the data after they re-identified one of the survey subjects—an acquaintance of one of the students.<sup>28</sup>

---

<sup>21</sup> *Id.*

<sup>22</sup> *Id.*

<sup>23</sup> Aloni Cohen, Sasho Nikolov, Zachary Schutzman & Jonathan Ullman, *Reconstruction Attacks in Practice*, DifferentialPrivacy.org (Oct. 27, 2020), <https://differentialprivacy.org/diffix-attack/>.

<sup>24</sup> Amitai Ziv, *Israel's 'Anonymous' Statistics Surveys Aren't So Anonymous*, Haaretz (Jan. 7, 2013), <https://www.haaretz.com/surveys-not-as-anonymous-as-respondents-think-1.5288950>.

<sup>25</sup> *Id.*

<sup>26</sup> *Id.*

<sup>27</sup> *Id.*

<sup>28</sup> *Id.*

These examples are not flukes, but evidence of a new and growing risk facing statistical agencies. The ability to reconstruct data records from overly accurate statistics is a mathematical certainty. A consensus study report published by the National Academies of Sciences, Engineering, and Medicine in 2017 concluded that traditional statistical disclosure methods “are increasingly susceptible to privacy breaches given the proliferation of external data sources and the availability of high-powered computing that could enable inferences about people or entities in a dataset, re-identification of specific people or entities, and even reconstruction of the original data.”<sup>29</sup> The research described above has borne this out.

#### **D. Reconstruction Attacks Enable Re-Identification Attacks.**

Reconstruction yields accurate records for a large swath of Census respondents, with names removed. Such data are often called “anonymized.” But “re-identification” of anonymized data is notoriously common and will be easier when leveraged by more modern datasets. Re-identification is the process of associating person-level data to the identities of actual people.<sup>30</sup> Once records are reconstructed, re-identification is relatively easy. It uses standard and well-known techniques and requires only access to commercial or public datasets that overlap with the anonymous records on some subset of the data fields. As the President’s Council of Advisors on Science and Technology wrote in 2014, “it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data.”<sup>31</sup>

---

<sup>29</sup> National Academies of Sciences, Engineering, and Medicine, Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps 104-05 (2017).

<sup>30</sup> Boris Lubarsky, *Re-identification of “Anonymized Data,”* 1 Geo. L. Tech. Rev. 202, 208-09 (2017).

<sup>31</sup> President's Council of Advisors on Science and Technology, *Report to the President, Big Data and Privacy: A Technology Perspective* (2014), #3205841v1

Examples of re-identification of anonymized health records abound. Sweeney re-identified patients in anonymized health records from Washington state using newspaper stories<sup>32</sup>; Yoo et. al. re-identified patients in Maine and Vermont using newspaper stories as well, even when such data is anonymized according to the principles set forth in the HIPAA Safe Harbor Standard<sup>33</sup>; and Sweeney et. al. re-identified individuals from the Northern California Household Exposure Study using the combination of tax data and online tools—such as a data broker website and Google Earth and Street View—even when the data were anonymized to standards beyond what is required by HIPAA’s Safe Harbor Standard.<sup>34</sup> In August 2016, the Australian Government released three billion records of billing data from its Medicare and Pharmaceutical Benefits Schemes, covering 10% of the Australian population (2.9 million people), anonymizing not only the patient but the medical provider as well.<sup>35</sup> Shortly after release, researchers re-identified all medical providers in the dataset.<sup>36</sup> Separately and without using re-identified provider information, the researchers manually re-identified at least five patients by linking approximate birth dates of children in the

---

[https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast\\_big\\_data\\_and\\_privacy\\_-\\_may\\_2014.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf).

<sup>32</sup> Latanya Sweeney, *Only You, Your Doctor, and Many Others May Know*, Technology Science (Sept. 2015).

<sup>33</sup> Yoo, Ji Su, Alexandra Thaler, Latanya Sweeney, & Jinyan Zang, *Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data*, Technology Science (Oct. 2018).

<sup>34</sup> Latanya Sweeney, Ji Su Yoo, Laura Perovich, Katherine E. Boronow, Phil Brown, & Julia Green Brody, *Re-identification Risks in HIPAA Safe Harbor Data*, Technology Science (Aug. 2017).

<sup>35</sup> Dr. Vanessa Teague, Dr. Chris Culnane, & Dr. Ben Rubenstein, *The Simple Process of Re-identifying Patients in Public Health Records*, Pursuit (Dec. 18. 2017), <https://pursuit.unimelb.edu.au/articles/the-simple-process-of-re-identifying-patients-in-public-health-records>.

<sup>36</sup> *See id.*

#3205841v1

dataset with publicly-available information on Wikipedia and news articles of public figure women such as politicians, athletes, and celebrities.<sup>37</sup>

In a 2018 study on privacy vulnerabilities in anonymized California Bar Exam data, Sweeney, von Loewenfeldt, and Perry were able to re-identify individuals in spite of the presence of four anonymization protocols put forth by ‘data privacy experts’ in the case of *Richard Sander et. al v. State Bar of California et. al* by utilizing a host of auxiliary information -- such as online graduation programs, attorney license date data, online alumni and club membership lists.<sup>38</sup> Also in 2018, the public transit authority of Victoria, Australia released two billion anonymized records of travelers in Melbourne.<sup>39</sup> Within three months researchers Culnane, Rubinstein & Teague had confirmed re-identifications of themselves, a co-traveller and a member of the Victorian State Parliament.<sup>40</sup>

At this point, re-identification of “anonymized” data is taken for granted by the academic privacy community. It is no longer an open research question.

#### **E. Reconstruction-Abetted Re-Identification Attacks Are a Realistic Threat.**

Reconstruction and re-identification require neither the skills nor resources of a government agency. The Census Bureau's reconstruction and re-identification demonstration used eighteen-year-old techniques and publicly available data; the attack used no confidential data

---

<sup>37</sup> *Id.*

<sup>38</sup> Latanya Sweeney, Michael von Loewenfeldt, & Melissa Perry, *Saying It's Anonymous Doesn't Make It So: Re-identifications of "Anonymized" Law School Data*, *Journal of Technology Science* (Nov. 12, 2018), <https://techscience.org/a/2018111301>.

<sup>39</sup> Josh Taylor, *Myki Data Release Breached Privacy Laws and Revealed Travel Histories, Including of Victorian MP*, *the Guardian* (Aug. 15, 2019), <https://www.theguardian.com/australia-news/2019/aug/15/myki-data-release-breached-privacy-laws-and-revealed-travel-histories-including-of-victorian-mp>.

<sup>40</sup> *Id.*

collected by the Bureau.<sup>41</sup> A journalism professor was able to reproduce the reconstruction attack in "about a week."<sup>42</sup> Reconstruction-abetted reidentification attacks could create risks to national security. Entities who possess substantial troves of non-public personal data about the U.S. population are particularly well positioned to perform re-identification attacks on reconstructed datasets. Such non-public data might be gathered in the ordinary course of business—by Google, Facebook, Twitter, and the many data brokers that legally profit from digital surveillance—and be used to advertise, influence, and silence.

Data breaches, such as the Office of Personnel Management (OPM) hack, are another major source of non-public data that could be used by an attacker. The OPM intrusion, widely attributed to hackers working with the Chinese government, exposed detailed files and security clearance background reports on more than 21.5 million individuals.<sup>43</sup> The files contained both relatively mundane data such as Social Security number, date and place of birth, and in the case of security clearance background reports extremely sensitive information including data about mental health, drug use and financial problems due to gambling.<sup>44</sup>

Reconstruction and re-identification attacks using data from the OPM breach or any of the many other data breaches occurring in the U.S. every year, by foreign governments and others with potentially adversarial interests could create risks to national security. For example, a foreign power could undermine confidence in the Census Bureau and depress future participation in the census by using Facebook or another social media platform to reveal to 50 million Americans that

---

<sup>41</sup> See Abowd Decl. ¶ 38, Doc. 41-1.

<sup>42</sup> Hansen, *supra* note 12.

<sup>43</sup> Kim Zetter, The Massive OPM Hack Actually Hit 21 Million People, *Wired* (July 9, 2015 4:25 PM), <https://www.wired.com/2015/07/massive-opm-hack-actually-affected-25-million/>.

<sup>44</sup> See *id.*



their data can be reconstructed and re-identified from census responses. This could be done selectively to target particular communities. For example, it could be targeted at a particular geographic area, such as zip code, and result in selectively depressing participation.

## **II. Census Confidentiality Protections Must Evolve to Address Today's Threats.**

### **A. Differential Privacy Is the Only Known Way to Protect Against Reconstruction Attacks.**

Differential Privacy is the only known method for protecting large-scale statistical releases against reconstruction attacks, and hence also against reconstruction-abetted re-identification attacks. Fifteen years after its invention, there is still no effective alternative to differential privacy for defending against this threat. Given the widely understood risks described above, and unique ability of differential privacy to address them, the Census Bureau's Data Stewardship Executive Policy Committee's (DSEP) decision that the Census Bureau should use differential privacy as the core of the 2020 Disclosure Avoidance System (DAS) is wise.<sup>45</sup>

In addition to being the only known approach available to protect large-scale statistical releases from reconstruction attacks, differential privacy has three properties that further advance the Census Bureau's twin mandates of providing useful statistical data and protecting the confidentiality of respondents: *First*, unlike all other technologies, differential privacy is future-proof. Commercial datasets are improved and created all the time. New attacks happen all the time: Sweeney galvanized re-identification; Dinur and Nissim discovered reconstruction. Future-proofing is particularly important given the number and type of statistics the Census Bureau publishes. *Second*, differential privacy is adversary agnostic, this means it will provide protection regardless of the motivation or financial, computational, and informational assets of the adversary.

---

<sup>45</sup> See generally Abowd Decl. ¶ 46, Doc. 41-1.

*Third*, differential privacy is measurable, allowing the public to quantify cumulative privacy loss as data are analyzed and re-analyzed, shared, and linked.<sup>46</sup>

Moreover, systems built on differential privacy do not require secrecy of the algorithm to protect confidentiality. The ability to publicly share the implementation choices in the 2020 DAS—which the Census Bureau has publicly committed itself to do—enables stakeholders, including policy makers, data subjects and data users, to assess the level of privacy protected through those choices.<sup>47</sup> This creates an unprecedented increase in transparency. Stakeholders will be able to review how the agency translates its obligation to produce useful statistical reports and protect the confidentiality of participants into technical design.

This means that with differential privacy, users of the data—for example the legislatures or participants in a Voting Rights Act case—can compute confidence intervals with confidence. It means that the Census Bureau can measure privacy loss over subsequent statistical releases and censuses. It means that whether the adversary is a hostile nation state or an angry teenager the confidentiality promises the Census Bureau makes will hold. Thus, differentially private systems allow the Census Bureau and other agencies that use them to, for the first time, measure and control the total privacy loss over the huge number of statistics and statistical products it releases.

---

<sup>46</sup> Cynthia Dwork, Frank McSherry, Kobbi Nissim, & Adam Smith, *Calibrating Noise to Sensitivity in Private Data Analysis*, *Journal of Privacy and Confidentiality* 17-51 (2016), <https://journalprivacyconfidentiality.org/index.php/jpc/article/download/405/388/>. See also Cynthia Dwork, Nitin Kohli & Deirdre Mulligan, *Differential Privacy in Practice: Expose Your Epsilons!*, 9 *Journal of Privacy and Confidentiality* (2019), <https://journalprivacyconfidentiality.org/index.php/jpc/article/view/689/685>.

<sup>47</sup> Abowd Decl. ¶ 62 (explaining that the Census Bureau has "committed to publicly releasing the entire production code base and full suite of implementation settings and parameters"), Doc. 41-1.

**B. The Census Bureau Cannot Tailor Its Confidentiality Protections to a Set of Predictable Risks as Suggested by Amicus Bambauer.**

As the reconstruction attacks described above reveal, it is impossible to figure out which attacks are likely and imprudent to assume that some whole category of attack is off the table. The Census Bureau has no crystal ball: they cannot know which attacks are likely to be real threats today, let alone over the 72-year time span during which they are obligated to protect confidentiality. It is not possible for the Census Bureau to assign probabilities to attacks. The Bureau does not know what motivates the attackers, or what information (other databases) they can access. The attacker could be Facebook (yes, the whole company; nothing in the attack is illegal), for example, or employees at Facebook (or another company with enormous troves of personal data) or a company that uses a Facebook application interface (as Cambridge Analytica did) to access the vast data sources available through a social media platform like Facebooks, or a malevolent individual or organization who scrapes personal data off the web. In addition, the attacks leveraged by the Census Bureau's internal researchers, and used in the other attacks described above, are only (some of) the attack methods we are aware of today. New attack methods can, and surely will, be designed by researchers and motivated attackers.

Privacy is a non-renewable resource. If the Census Bureau assumed that a specific attack was unlikely, failed to protect against it, but then found that this attack was, in fact, going on, there would simply be no means of re-asserting additional protection. Once data with specific reconstruction or reidentification vulnerabilities has been released, they cannot be withdrawn. Protecting against worst-case attacks is the easiest way of protecting against as yet unknown realistic attacks (or attacks that will become realistic at some point in the future). It is completely misleading to characterize protection against worst-case attacks as a preoccupation with highly

unrealistic attacks (e.g., attacker who knows all but one record). The attack carried out by the Census Bureau on the 2010 release is *exactly* a case in point: it was, at the time, a new attack method that defeated the 2010 DAS protections. It was easy to carry out. No genius will be required to carry out a similar attack. In the foreseeable future, someone will likely publish or market a script or code to show others how to replicate the attack—turning yesterday’s innovation into tomorrow’s readily accessible weapon, usable by anyone who can download it.

### **C. Heuristic Alternatives Have Several Limitations.**

*First*, they are evaluated based on outside data sources and algorithmic techniques that are available at the time. *By definition they are not “future-proof.”* The Census’ own internal attacks on the 2010 DAS demonstrate how fragile the guarantees can be. Consequently, such heuristics are just not reliable.

*Second*, heuristic approaches do not provide a measure of privacy loss. To the extent they “work”—which is typically unproven and indeed unprovable—they rely in part on secrecy. Statistical agencies are reluctant to be fully transparent about the techniques they use because adversaries can use such information to build attacks. Yet without detailed knowledge of the heuristics, it is impossible for users of the data—legislatures or participants in a Voting Rights Act case, for example—to evaluate how certain their conclusions are, by for example, computing confidence intervals.

### **III. Distinguishing the Census Bureau's 2020 Disclosure Avoidance System (2020 DAS) and Differential Privacy.**

"Differential privacy" is a mathematical definition that some algorithms satisfy and others do not. Algorithms that satisfy the definition are called "differentially private." There are many procedures (or algorithms) that “satisfy”—meaning “adhere to”—differential privacy. For

instance, some differentially private algorithms operate by perturbing individual fields in data records, while others inject noise into the outcome of a computation or even into carefully chosen intermediate steps. A useful metaphor is to think of differential privacy as a security or safety standard that can be met in several different ways: a stopping distance standard for car brakes, for example, does not specify how the brakes should operate, but simply how quickly the vehicle must come to a stop.

Differentially private algorithms also come with a privacy budget that limits how much their output can help an attacker to make inferences about individuals in the data set. Even for a given budget, there are many different algorithms that satisfy the standard. Some will be far more accurate than others. Without familiarity with the data, it is generally difficult to determine the most accurate differentially private algorithm for a particular desired task and with a particular privacy loss budget. These kinds of questions are the subject of much research in the field.

Given this, it is crucial to distinguish discussions (critical or not) of differential privacy from discussions of the accuracy of particular algorithms or implementations. Many of the documents submitted to the court as part of this case use the terminology in confusing ways, mistaking "differential privacy" for the current proposed implementation (the proposed 2020 DAS). These include the plaintiff's original brief as well as the expert report by Dr. Barber and the amicus brief by Jane Bambauer.

#### **IV. The 2020 DAS Does Not Use Statistical Inference.**

The briefs and other materials at times use technical terms without precision to argue about the construction and application of statutory definitions. We do not offer an opinion on the correct interpretation of the statute. However, we do wish to clarify for the Court the point at which differential privacy is part of the Census Bureau's workflow and the meaning of several terms of

#3205841v1

art in statistics, and specifically differential privacy, that are used imprecisely, and at times inaccurately, within the record.

The DAS is applied after enumeration is completed. The DAS consists of two steps: step one uses differential privacy, and step two is a post-processing step that relies on optimization tools. Neither step involves “statistical inference” as defined in the relevant fields.

**A. Differential Privacy As Used In the 2020 DAS Does Not Use Statistical Inference.**

The introduction of carefully calibrated privacy-infusing random noise carried out in the first step of the 2020 DAS is most accurately viewed as “fuzzing” the details, much as faces of bystanders may be intentionally blurred out in pictures or videos. In this sense, the techniques used in the first step of the 2020 DAS are simply a more mathematically rigorous and principled alternative to the swapping than was done in the 2010 DAS, which also “fuzzed” details, typically by exchanging minority households with majority households, and which resulted in more apparent homogeneity than was actually the case. Abowd describes swapping as a form of “noise infusion.”<sup>48</sup> No inferences are drawn, statistical or otherwise.<sup>49</sup>

Plaintiff’s claim that “differential privacy is . . . an unlawful ‘statistical method’” and that “[i]t is clear that differential privacy falls into this category” is inaccurate.<sup>50</sup> Barber’s expert declaration, which Plaintiffs quote as support for their claims, belies their argument that differential privacy is a statistical method:

---

<sup>48</sup> Abowd Decl. ¶ 24, Doc. 41-1.

<sup>49</sup> Statistical inference is a term of art. See the definition given by Sir R. D. Cox (Oxford), inaugural winner of the International Prize in Statistics, in Appendix B.

<sup>50</sup> Pl.’s Motion for a Prelim. Injunction, Pt. for a Writ of Mandamus, and Mem. in Support (Mar. 11, 2021) at 38, Doc. 3.

Privacy is introduced into the data by introducing random error through sampling from statistical distributions with parameters set to a desired level of variance . . . *Differential privacy is thus an application of statistical processes and methods to adjust the original counts of the census to protect the privacy of individual records.*<sup>51</sup>

Plaintiff is correct that the fuzzing in the 2020 DAS involves sampling from statistical distributions—usually called probability distributions. But this is not statistical inference.

**B. The Post-Processing in Step Two of the DAS Does Not Use Statistical Inference.**

The second step of the 2020 DAS modifies the privacy-infused statistics to satisfy certain constraints, such as consistency with state enumeration totals and certain publicly known information including the total number of housing units at the Census block level, the total number of group housing units by type in each block, and to ensure non-negative counts. The simplest analogy is to round a number to the nearest integer, for example, rounding 12.2 to 12, or 16.8 to 17 (in this rounding example, the “constraint” is that values reported are whole numbers).

Once again, there is no inference, statistical or otherwise, in this step. Michael Hawes’ presentation is mistaken in using this term.<sup>52</sup> Hawes was referring to the use of L2 optimization, used in the 2020 DAS to perform step two as described above. Although L2 optimization can be used in statistical inference, that is not its purpose in the 2020 DAS. Plaintiff and their expert mistakenly rely on Hawes’ misuse of the term.<sup>53</sup> The authoritative source on the approach underlying the second step in the 2020 DAS extensively describes the problem the L2 optimizer

---

<sup>51</sup> First Decl. of Dr. Michael Barber at 16-17, Doc. 3-5.

<sup>52</sup> Michael Hawes, *Differential Privacy and the 2020 Decennial Census*, U.S. Census Bureau at slide 40 (Jan. 28, 2020), [https://zenodo.org/record/4122103/files/Privacy\\_webinar\\_1-28-2020.pdf](https://zenodo.org/record/4122103/files/Privacy_webinar_1-28-2020.pdf).

<sup>53</sup> Pl.’s Reply in Support of Their Request for the Appointment of a Three-Judge Court (Mar. 25, 2021) at 1, Doc. 25; Second Expert Report of Dr. Michael Barber at 12, Doc. 25-2.

is used to address which, as described above, has nothing to do with inference – the Census Bureau already knows the confidential data!<sup>54</sup>

Neither step of the 2020 DAS satisfies the definition of statistical inference as laid out by Sir R. D. Cox (see Appendix) and understood in the field.

### CONCLUSION

The Census Bureau is tasked with providing myriad useful aggregate statistics and protecting the confidentiality of respondents. As all statistics computed from a dataset reveal small hints about the individual data records, reconstruction attacks make the Census Bureau’s task more challenging. The 2010 DAS used traditional disclosure avoidance techniques that have not aged well. The Census Bureau’s research, and the other well-known reconstruction attacks, document the inability of those approaches to provide any meaningful level of confidentiality today. The Census Bureau—like other statistical agencies—must adopt protections to fit the changing threats. Thanks to fifteen years of research on differential privacy, the Census Bureau has the tools to meet its statutory obligation to both provide useful statistical data and provide future-proof protection of privacy. These advances have allowed the Census Bureau to—for the very first time—measure privacy loss, fully disclose the way in which the DAS protects confidentiality, permit the computation of confidence intervals, and advance public debate about the balance between privacy and accuracy.

---

<sup>54</sup> John Abowd et al., *Census TopDown: Differentially Private Data, Incremental Schemas, and Consistency with Public Knowledge*, U.S. Census Bureau at 6 (2019), <https://columbia.github.io/private-systems-class/papers/Abowd2019Census.pdf>.



Respectfully submitted this 29th day of April, 2021.

/s/ Michael B. Jones

Michael B. Jones

Admitted Pro Hac Vice

Georgia Bar No. 721264

**BONDURANT, MIXSON &  
ELMORE, LLP**

1201 W. Peachtree Street, NW

Suite 3900

Atlanta, GA 30309

Telephone: (404) 881-4100

Facsimile: (404) 881-4111

Email: [jones@bmelaw.com](mailto:jones@bmelaw.com)

Shannon L. Holliday (ASB-5440-Y77S)

Copeland, Franco, Screws & Gill, P.A.

P.O. Box 347

Montgomery, AL 36101-0347

Telephone: (334) 834-1180

Facsimile: (334) 834-3172

Email: [holliday@copelandfranco.com](mailto:holliday@copelandfranco.com)

***Counsel for the Data Privacy Experts***

**UNITED STATES DISTRICT COURT  
FOR THE MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION**

THE STATE OF ALABAMA, <i>et al.</i> ,	)	
	)	
Plaintiffs,	)	
	)	
v.	)	Civil Action No.
	)	3:21-CV-211-RAH
UNITED STATES DEPARTMENT OF	)	
COMMERCE, <i>et al.</i> ,	)	
	)	
Defendants.	)	

**CERTIFICATE OF SERVICE**

I hereby certify that on the 29th day of April, 2021, I electronically filed the foregoing **AMICUS BRIEF OF DATA PRIVACY EXPERTS** with the Clerk of Court using the CM-ECF system which will automatically send e-mail notification of such filing to all parties of record.

/s/ Michael B. Jones  
Michael B. Jones  
(Ga. Bar No. 721264)

## APPENDIX A – LIST OF AMICI CURIAE

*Institutions are listed for affiliation purposes only. All signatories are participating in their individual capacity, not on behalf of their institutions.*

- Ryan Calo
  - Lane Powell and D. Wayne Gittinger Professor
  - University of Washington School of Law
- Ran Canetti
  - Professor of Computing and Data Science
  - Head of the Center for Reliable Information Systems and Cyber Security
  - Boston University
- Aloni Cohen
  - Postdoctoral Associate, Hariri Institute for Computing and the School of Law
  - Boston University
- Cynthia Dwork<sup>†</sup>
  - Gordon McKay Professor of Computer Science and Applied Mathematics and Radcliffe Alumnae Professor
  - Harvard University
  - Distinguished Scientist, Microsoft Research
  - Harvard Co-Principal Investigator on Cooperative Agreement CB16ADR0160001 from the U.S. Census Bureau to Georgetown University, and Cooperative Agreement CB20ADR0160001 from the U.S. Census Bureau to Boston University
- Roxana Geambasu
  - Associate Professor of Computer Science
  - Columbia University
- Somesh Jha
  - Lubar Professor of Computer Sciences
  - University of Wisconsin, Madison
- Nitin Kohli
  - PhD Candidate
  - UC Berkeley School of Information
- Aleksandra Korolova

---

<sup>†</sup> These specific Amici have current research funding from the Census Bureau (Cooperative Agreements CB16ADR0160001 and CB20ADR0160001) to work on differential privacy generally. This brief is submitted wholly independently and not reliant on any funding or non-public information provided by the Census Bureau. In particular, none of the authors participated in the implementation of the 2020 Disclosure Avoidance System. The opinions, findings, conclusions and recommendations expressed herein are those of the authors and do not necessarily reflect the views of their employers or the organizations with which they collaborate.

- WiSE Gabilan Assistant Professor of Computer Science
  - University of Southern California
- Jing Lei
  - Associate Professor of Statistics and Data Science
  - Carnegie Mellon University
- Katrina Ligett
  - Associate Professor of Computer Science
  - The Hebrew University of Jerusalem
- Deirdre K. Mulligan
  - Professor
  - UC Berkeley School of Information
- Omer Reingold
  - Professor of Computer Science
  - Stanford University
- Aaron Roth
  - Professor of Computer and Information Sciences
  - University of Pennsylvania
- Guy N. Rothblum
  - Associate Professor of Computer Science and Applied Mathematics
  - Weizmann Institute of Science
- Benjamin Rubinstein (*Inadvertently omitted from Motion for Leave to File Amicus Brief.*)
  - Professor of Computing and Information Systems
  - Associate Dean (Research), Faculty of Engineering and Information Technology
  - The University of Melbourne, Australia
- Aleksandra (Seša) Slavkovic<sup>‡</sup>
  - Professor, Departments of Statistics and Public Health Sciences
  - Associate Dean for Graduate Education, Eberly College of Science
  - The Pennsylvania State University
- Adam Smith<sup>†</sup>
  - Professor of Computer Science and Electrical and Computer Engineering
  - Boston University
  - Co-Principal Investigator on Cooperative Agreement CB16ADR0160001 from the U.S. Census Bureau to Georgetown University, and Cooperative Agreement CB20ADR0160001 from the U.S. Census Bureau to Boston University.
- Kunal Talwar<sup>§</sup>
  - Senior research scientist, Apple

---

<sup>‡</sup> Member, Committee on National Statistics CNSTAT Census DAS Expert Group on noisy measurements

<sup>§</sup> Census Scientific Advisory Committee

- Salil Vadhan<sup>†\*\*</sup>
  - Vicky Joseph Professor of Computer Science and Applied Mathematics
  - Harvard University
  - Harvard Principal Investigator on Cooperative Agreement CB16ADR0160001 from the U.S. Census Bureau to Georgetown University, and Cooperative Agreement CB20ADR0160001 from the U.S. Census Bureau to Boston University
- Larry Wasserman
  - UPMC Professor of Statistics and Data Science
  - Carnegie Mellon University
- Daniel J. Weitzner
  - 3Com Founders Principal Research Scientist
  - Computer Science and Artificial Intelligence Laboratory
  - Massachusetts Institute of Technology

---

<sup>\*\*</sup> Member of the CNSTAT Census DAS expert group on post processing

## APPENDIX B

- The Oxford Dictionary defines statistical inference as "The theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling." Statistical Inference, Oxford English Dictionary, [https://www.lexico.com/en/definition/statistical\\_inference](https://www.lexico.com/en/definition/statistical_inference).
- The authoritative scholarly reference is D. R. Cox, *Some Problems Connected with Statistical Inference*, 29 Ann. Math. Statist, 357, 357 (1958):
  - "A statistical inference...[is] a statement about statistical populations made from given observations with measured uncertainty. An inference in general is an uncertain conclusion. Two things mark out statistical inferences. First, the information on which they are based is statistical, i.e. consists of observations subject to random fluctuations. Secondly, we explicitly recognise that our conclusion is uncertain, and attempt to measure, as objectively as possible, the uncertainty involved. Fisher uses the expression 'the rigorous measurement of uncertainty.'"
  - "A statistical inference carries us from observations to conclusions about the populations sampled. A scientific inference in the broader sense is usually concerned with arguing from descriptive facts about populations to some deeper understanding of the system under investigation."

IN THE UNITED STATES DISTRICT COURT  
FOR THE MIDDLE DISTRICT OF ALABAMA  
EASTERN DIVISION

STATE OF ALABAMA, *et al.*,

Plaintiffs,

v.

UNITED STATES DEPARTMENT OF  
COMMERCE, *et al.*,

Defendants.

Case No. 3:21-CV-211-RAH-ECM-KCN

SUPPLEMENTAL DECLARATION OF JOHN M. ABOWD

I, John M. Abowd, make this supplemental Declaration pursuant to 28 U.S.C. § 1746, and declare that under penalty of perjury the following is true and correct to the best of my knowledge. I am submitting this Declaration to supplement the Declaration I submitted in this case on April 13, 2021. In this supplemental Declaration, I clarify and respond to allegations and claims made by plaintiffs and their declarants.

**REVERTING TO SWAPPING WILL FURTHER DELAY THE RELEASE OF REDISTRICTING DATA.**

1. In my prior Declaration, I stated that there would be substantial additional delays to the release of the redistricting data were the Court to require the Census Bureau to revert to using swapping for the 2020 Census (Abowd Decl., ¶¶ 84–86). Plaintiffs counter, “there’s good reason to think the 2010 methods could be applied more quickly than still-in-development differential privacy.” (Reply, p. 21). This is categorically false. Given the scale, complexity and critical importance of the decennial census, the Census Bureau has developed and consistently applied rigorous standard operating procedures to ensure the integrity and reliability of census data processing. Reverting to swapping for the 2020 Census would require numerous analysis, policy, auditing, deployment, system testing and quality assurance steps before swapping could be used in the 2020 Census production workflow.
2. Some steps that would be required include that the Census Bureau’s Disclosure Review Board (DRB) would need to assess the disclosure risks and develop the proposed swapping algorithm, parameters and swap rates to be used. Next, the DRB would need to assess and document the residual risk of disclosure for this methodology. The Data Stewardship Executive Policy (DSEP) Committee would then need to review and approve the DRB’s proposal and residual risk assessment. Once approved, technical staff would need to program the swapping algorithms for use in the 2020 Census computing environment and their work would need to be audited to ensure that the software accurately implements the selected swapping rules, parameters and rates. Once



audited, the swapping software could be deployed to the 2020 Census computing environment, where it would have to undergo extensive, mandatory system integration testing before it could be used in production. Even if all these steps were expedited, this whole process could take 24-28 weeks<sup>1</sup> and would inevitably delay further the release of the redistricting data product.<sup>2</sup>

3. The 2020 Disclosure Avoidance System using the TopDown Algorithm (TDA) is fully operational and has already completed all necessary auditing and system integration testing currently required for 2020 Census Information Technology systems. All that remains is the final Operational Readiness Review on May 20, 2021 and the final setting and allocation of the privacy-loss budget by DSEP in June, incorporating data user feedback from the April 2021 demonstration data. Under any scenario, using the 2020 DAS will enable the Census Bureau to release the redistricting data sooner than would be possible if we were required to revert to swapping.

---

<sup>1</sup> The estimate is based on the time required to completely refactor the code base for swapping, port the refactored code to the production environment for the 2020 Census, repeat the Test Readiness, Production Readiness and Operational Readiness Reviews with the same protocols used for the current DAS, then resume the production sequence to produce a clean, certified Microdata Detail File.

<sup>2</sup> Just like in prior years, the disclosure avoidance needs to be applied prior to the release of the redistricting data, or any data product other than apportionment. The federal government and the broader statistical disclosure limitation field have long acknowledged the necessity of considering all releases of related data when making decisions regarding disclosure risk. Office of Federal Statistical Policy and Standards (1978) Statistical Policy Working Paper #2 “Report on Statistical Disclosure and Disclosure Avoidance Techniques” p. 14, available at <https://nces.ed.gov/FCSM/pdf/spwp2.pdf#:~:text=Policy%20and%20Standards%20Statistical%20Policy%20Working%20Paper%202,Economist%20Office%20of%20Federal%20Statistical%20Policy%20and%20Standards>. See also: Cox (1976) and Fellegi (1972).

## THE CENSUS BUREAU'S TOPDOWN ALGORITHM CAN BE TUNED TO MAKE POPULATION COUNTS EFFECTIVELY INVARIANT

4. The privacy accounting framework of differential privacy and the hundreds of finely tunable parameters of the Census Bureau's 2020 TDA are extremely nimble and precise. For example, the Census Bureau could meet numerical accuracy targets for block-level population counts through reallocation of privacy-loss budgets using the tuning parameters. Allocating a sufficiently high privacy-loss budget for population counts at the block level could result in nearly all block population counts being reported exactly as enumerated. Note, however, that the Census Bureau has already leveraged the flexibility and precision of the TDA to meet the accuracy targets we established for the redistricting and Voting Rights Act use cases. As I explained in my prior Declaration, keeping the block-level population variant is important because even the slightest uncertainty in each block-level population provides exponential protection to the data set as a whole. We could reallocate privacy-loss budget to meet accuracy requirements within the current algorithm and schedule, but that would be at the expense of accuracy for other characteristics.<sup>3</sup>

## THE POTENTIAL CONSEQUENCES OF HOLDING BLOCK-LEVEL POPULATION INVARIANT.

5. Holding block-level populations invariant would present a number of challenges and would be difficult to implement within our existing Disclosure Avoidance System using TDA.<sup>4</sup> First, implementing this invariant would risk further delaying the Redistricting Data. A new invariant would also put at risk the fitness-for-use of the

---

<sup>3</sup> Reallocating privacy-loss budget for block-level population accuracy implies that other data such as voting age, race, ethnicity, sex, and age will be less accurate.

<sup>4</sup> The ability of DAS to find a feasible Microdata Detail File in the presence of an additional invariant, e.g., block-level population totals, depends upon proving that a technical condition in mixed integer linear programming remains true in the presence of the new invariant. That condition has not been verified for the production version of the DAS

remaining 2020 Census data because it would remove *all* confidentiality protection on a key identifier used in re-identification attacks – the census block. If the Census Bureau removes that uncertainty by forcing a block-level population invariant, stronger disclosure avoidance (more noise) would have to be used for other variables like sex, age, race, and ethnicity.

6. Requiring the Census Bureau to hold block-level population invariant could further delay the release of the Redistricting Data. The 2020 Disclosure Avoidance System using TDA can hold certain tabulations invariant. Current invariants are already programmed into the algorithm including total population counts at the state level, the number of housing units at the block level, and the number and major type of group quarters facilities at the block level.
7. The 2020 DAS TopDown Algorithm performs a series of complex optimizations at each geographic level, from the nation down to individual census blocks. Imposing constraints on these optimizations through invariants limits the number of possible solutions to these optimization steps and runs the risk of the algorithm either taking longer than expected to complete the optimization or crashing entirely.<sup>5</sup>
8. The Census Bureau has extensively tested the stability and performance of the 2020 DAS with the current list of invariants. Unless similar testing and analysis is done on the potential impact that including this additional invariant might have, we cannot guarantee that the DAS will be able to complete its production run of the 2020 Census

---

planned for the P.L. 94-171 Redistricting Data Summary File in the presence of a block total population invariant.

<sup>5</sup> The operational vulnerability caused by invariants is not unique to our implementation, nor to differential privacy as a whole; swapping algorithms like those used in 2010 could face similar “unsolvable” situations if those algorithms are unable to find households in the target geographies that match on the key swapping characteristics (i.e., the invariants).

redistricting data in the period of time currently allotted for disclosure avoidance processing in the production schedule.

9. To add an invariant for total population counts at the block level to the algorithm, we would need to modify the algorithm – in other words, we need to complete a mathematical analysis of the full equation system that produces the Microdata Detail File to ensure that a mathematical solution to the system exists and can be found in finite time using the state-of-the-art commercial optimization software embedded in the DAS. The production code base is scheduled for finalization in its Operational Readiness Review (ORR) on May 20, 2021.<sup>6</sup>
10. There are many data processing steps that need to occur between the release of the apportionment data and the release of the redistricting data product. The application of disclosure avoidance is one small, and relatively short, step in that process. Under our current production schedule, targeting release of the redistricting data in the legacy file format by August 16, 2021, each sequential data processing stage has a tightly constrained duration and there is no margin for schedule slippage. If any processing step takes longer than it is allocated, the release date would likely be delayed. The introduction of a new, untested invariant poses just such a risk.
11. Including a block-level population count invariant would also impact data quality for the remaining 2020 Census data. In my prior Declaration, I explained how including even minimal amounts of uncertainty in block level population counts greatly reduces the ability of an attacker to perform a reconstruction-abetted re-identification attack (Abowd Decl., ¶42). Were the Census Bureau to impose a block-level invariant on population counts, we would necessarily need to apply additional privacy protections

---

<sup>6</sup> After the Operational Readiness Review, the privacy-loss budget can still be adjusted and reassigned, but the optimization problems the DAS solves are locked.

to tabulations of other characteristics (sex, age, race, and ethnicity) to meet our obligations under 13 U.S. Code §§ 8(b) and 9. These additional protections, either in the form of table suppression or reduced privacy-loss budget would have deleterious consequences for redistricting and Voting Rights Act enforcement, as well as other statutory uses of decennial Census data, including the Population Estimates program and federal funding allocations. If the goal is to achieve down-to-the-person accuracy on block-level populations, then the correct way to accomplish this is to change the allocation of privacy-loss for that variable.

**PLAINTIFFS MISCONSTRUE HOW THE TDA PROCESSES TRIBAL AREAS.**

12. The 2020 Census DAS TopDown Algorithm operates along a geographic hierarchy; this is what ensures that the accuracy of statistics increases as the underlying population increases. The standard hierarchy starts at the national level, then descends to state, county, tract, block group, and finishes at the individual census blocks. This hierarchy posed some challenges early in the development of the DAS because it was difficult to ensure high levels of accuracy for geographic entities that split these geographic levels (e.g., for incorporated places that contain portions of different census tracts). The Census Bureau addressed this challenge by implementing several changes to the geographic hierarchy used by the TDA to improve accuracy for all “off-spine” geographic entities, like voting districts. One change to the geographic hierarchy that we implemented in September 2020 was to create an alternate geographic processing hierarchy for federally recognized American Indian and Alaska Native (AIAN) tribal areas within each state.

13. Plaintiffs allege that the separation of the AIAN tribal areas within the geographic hierarchy demonstrates that the Census Bureau is “prioritizing the accuracy of the data for certain racial and ethnic groups over others” (Reply, p.39). This is false. The changes to the AIAN geographic hierarchy were implemented to address the distinct

legal and political status of those geographies and to reflect the government-to-government relationship we have with federally recognized tribes. County, tract, block group and block statistics for these AIAN tribal areas receive the same allocation of privacy-loss budget (level of accuracy) as their corresponding geographies in the remainder of their states.<sup>7</sup> In fact, the isolation of the AIAN tribal areas is very similar to the way the TDA post-processing isolates group quarters facilities at the block group level from their surrounding non-group quarters populations. Privacy-loss budget, and its corresponding impact on accuracy, is allocated by data table element. Within each such element, all demographic sub-groups receive the same allocation of privacy-loss budget. Thus, the accuracy improvements derived from privacy-loss budget allocation apply to all demographic groups equally. The TDA does not, and will not, allocate greater privacy-loss budget to any particular demographic group or subgroup over another.

**THE 2000 DEPARTMENT OF JUSTICE LETTER.**

14. Plaintiffs quote two Census Bureau sources that referenced a 2000 “agreement” between the Census Bureau and the U.S. Department of Justice that supposedly established block-level invariants as a legal requirement (Reply, p.10). I personally investigated the history of this supposed agreement after it was raised by staff within the Census Bureau. After diligent research, we found that the supposed “agreement” was a March 15, 2000 letter from Acting Assistant Attorney General Bill Lee, of the DOJ Civil Rights Division, to Census Bureau Director Kenneth Prewitt, merely stating that the attorneys at the Civil Rights Division did “not believe that the application of [the proposed] disclosure avoidance techniques [for the 2000 Census] will impair the

---

<sup>7</sup> As I said in my first Declaration, accuracy is measured as the absolute error in the counts, not relative percentage errors.

use of these data for enforcement of civil rights programs.” The body of the letter is below and the full letter is attached as an Appendix.

Dear Dr. Prewitt:

This is in regard to the Census 2000 Redistricting Data and confirms the February 25, 2000 telephone conversation between Marshall Turner of your staff and my Deputy, Anita Hodgkiss.

At a November 29, 1999 meeting here at the Civil Rights Division, you and your staff discussed with Ms. Hodgkiss the possible need to apply certain disclosure avoidance techniques to the detailed race and ethnicity statistics included in the Census 2000 Redistricting Data files in order to meet the confidentiality requirements of Title 13, U.S.C. We have reviewed the information you provided at the November 29 meeting and we do not believe that the application of these disclosure avoidance techniques will impair the use of these data for enforcement of civil rights programs.

We greatly appreciate your kind assistance.

Sincerely,

Bill Lann Lee  
Acting Assistant  
Attorney General  
Civil Rights Division

15. Over the subsequent years, misinterpretation and faulty recollection of the content of this letter by Census Bureau staff led to the perpetuation of an erroneous oral history on this subject. As the Census Bureau began developing the requirements for the 2020 DAS, we retrieved and reviewed the original letter. As can be seen from the face of the letter, it contains no agreement or legal analysis requiring *any* invariants, let alone block-level population.

#### THE CENSUS BUREAU’S CONCERN ABOUT LOCATION PROTECTION IS NOT NEW

16. Plaintiffs assert that the Census Bureau’s decision not to hold population counts invariant at the block level, in order to enhance respondents’ location protection, reflects

a “new interpretation” by the Census Bureau regarding its obligations to protect confidentiality (Reply, p. 29). To the contrary, the Census Bureau has long recognized the growing disclosure risk of releasing highly accurate data for small geographic units. That was precisely the rationale behind the table suppression methodologies used prior to 1990 and the transition to swapping for the 1990, 2000 and 2010 Censuses.

17. Accurate and precise location information significantly increases the risk of re-identification (making it easier to match individuals to addresses contained in external data) because of the prevalence of persons in the population who have unique values for the combination of census block, age (in years) and sex. The Census Bureau has long employed disclosure avoidance methods to reduce this risk. Geographic aggregation (reporting only at higher levels of geography) was the primary protection afforded by table suppression in 1970 and 1980 and continued to be used through 2010 for the 2010 Census Public-Use Microdata Sample. Swapping, as used for the 1990-2010 Censuses, sought to further counter this growing risk by attempting to protect the individuals considered most vulnerable when reporting highly accurate block-level statistics. Between 1990 and 2010, the swapping methodologies and swap rates used evolved in an attempt to keep pace with the growing risks of releasing highly accurate information at the block level. But, as I referenced in my prior Declaration, the Census Bureau recognized even prior to the publication of 2010 Census data that the risks of publishing highly accurate block-level statistics were continuing to increase, and would need to be further evaluated in the context of 2020 Census planning (Abowd Decl. ¶37, fn33). Our adoption of differential privacy and our removal of the invariant on population counts at the block level are a highly effective mechanism for countering this growing threat of re-identification, while continuing to produce high quality statistics about the nation.



## **DR. STEVEN RUGGLES IS NOT AN EXPERT IN DIFFERENTIAL PRIVACY**

18. I am familiar with the work of Dr. Ruggles. He is one of the world's leading experts on and proponents of the use of household microdata to advance social science with demographic modeling. I have collaborated with him on multiple large-scale projects, including the dissemination of the demonstration data associated with the TopDown Algorithm. In my opinion, however, he does not have significant experience with the capabilities of modern statistical disclosure limitation based on the principles of differential privacy to render an expert opinion about this matter, specifically on the subject of the relation between disclosure limitation methods and the underlying mathematical theory of differential privacy for privacy-loss accounting. While Dr. Ruggles has published some articles on differential privacy, all are merely critiques of the Census Bureau's 2020 Disclosure Avoidance System that contain the same types of errors as the errors in the report submitted in this case. In addition, Dr. Ruggles conflates early iterations of the DAS released for demonstration purposes with its capabilities in production form.
19. One simple error permeates his report: Dr. Ruggles confuses the concepts of privacy-loss accounting using differential privacy with the specific disclosure limitation technique the Census Bureau plans to use, the TopDown Algorithm.
20. Differential privacy is not an algorithm or a disclosure limitation technique – it is an accounting method for evaluating and comparing risks from different disclosure limitation techniques. One way to think of differential privacy is like the accounting methods used by businesses to track expenditures and identify waste of resources. Differential privacy quantifies the privacy loss from making certain data public, and

it quantifies the privacy protection provided by applying different statistical disclosure limitation methods. An organization using differential privacy can compare different disclosure limitation methods and detect and fix vulnerabilities that could lead to significant privacy loss. The specific disclosure avoidance technique that the Census Bureau plans to use is called the TopDown Algorithm – it is not called “differential privacy.” Differential privacy is used to measure the disclosure risk after the Census Bureau applies its TopDown Algorithm (TDA).

21. Differential privacy can be used with a variety of statistical disclosure limitation techniques. Traditional, and very old, statistical disclosure limitation techniques such as randomized response (Warner 1965) and noise infusion (Evans et al. 1998) are often used with differential privacy accounting. But modern research has designed more efficient new techniques that improve accuracy for the same amount of privacy loss. For example, Google and Apple have used randomized response (a technique invented in 1965) combined with differential privacy accounting and accuracy improvements to collect mobile device usage information (citations in my original Declaration), while protecting user identities and activities on their phones.
22. Rather than being an “entirely new approach,” the TopDown Algorithm is an improvement over traditional methods based on new ideas that result from scientific research. To create TDA, the Census Bureau used differential privacy methods to improve the efficiency of noise infusion compared to its traditional data swapping approach.
23. Disclosure risk assessments, such as those cited by Dr. Ruggles, are known to underestimate true disclosure risk – these assessments can only measure the success of the specific privacy attack that they consider. As a result, if these assessments estimate high disclosure risk, then the true disclosure risk is high, but if these assessments estimate low disclosure risk, then no conclusions can be drawn. These attack-specific methods also ignore a myriad of other feasible privacy attacks. Methods focused on

a specific attack strategy are not capable of measuring disclosure risk more generally, let alone the disclosure risk possible after continued developments in computer hardware as well as improvements in the algorithms used in these privacy attacks. For this reason, such attack-specific methods cannot be used as the sole measures of disclosure risk.

24. Differential privacy-loss accounting is necessary because older disclosure limitation methods are less reliable at estimating disclosure risk. Often they severely underestimate disclosure risk, as shown by a successful reconstruction attack on Aircloak's Difix system by Cohen and Nissim in (2020, 2018) that used the same linear programming methods as the Census Bureau used to perform its simulated reconstruction-abetted re-identification attack on the 2010 Census. To re-iterate, prior to differential privacy, there was no satisfactory method for tracking privacy loss. Prior methods were based on assumptions about the attacker's information and technology but could severely underestimate the risk if those assumptions were wrong.
25. Dr. Ruggles' statement "[i]t has long been recognized, however, that there is no direct relationship between the level of  $\epsilon$  and the risk of disclosing identities" is incorrect and his reference to McClure and Reiter (2012) misconstrues their result. The epsilon parameter in differential privacy, when accurately measured, is directly related to disclosure risk (Wasserman and Zhou, 2010); specifically it limits the statistical power of all possible tests for whether a particular individual's data record (or portions thereof) was used to produce a collection of statistics versus the record of another, arbitrary individual. This is exactly the same identity disclosure definition used by McClure and Reiter.<sup>8</sup> The latter show, contrary to Dr. Ruggles' claim, that some attack models do not succeed even when the differential privacy parameter  $\epsilon$  is large. This means

---

<sup>8</sup> Technically, both methods set up the statistical re-identification hypothesis such that the likelihood ratio, the contribution of the data to the attacker's inference about re-identification, is the same.

the data may be safe from that particular attack, not that re-identification and epsilon are unrelated. Wasserman and Zhou show that *any* identity attack model is limited by  $\epsilon$  because it constrains the optimal test statistic for an identity disclosure whereas McClure and Reiter focus on very specific attacks. The body of work supporting differential privacy shows that the larger  $\epsilon$  is, the more likely some identity attack model will succeed, but even large values of  $\epsilon$  can effectively protect against specific attack models, when  $\epsilon$  is allocated strategically as demonstrated by McClure and Reiter.

**DR. RUGGLES' DISCLOSURE RISK ASSESSMENT IS FLAWED AND UNDERESTIMATES THE ACTUAL RISK FROM A RECONSTRUCTION-ABETTED RE-IDENTIFICATION ATTACK.**

26. Dr. Ruggles has mischaracterized the risk from reconstructed microdata for the entire population of the 2010 Census. His own risk assessment of the Census Bureau's 2010 data release is flawed, even using the standards of the statistical disclosure limitation literature that pre-dated the invention of privacy-loss accounting methods like differential privacy.

27. Since the influential work of Duncan and Lambert (1989) the risk of identity disclosure for a microdata record has been measured by the probability that the record is a population unique on key variables that can be used for record linkage to external data. As I defined in my first Declaration, population uniques have a combination of key characteristics that occurs exactly once in the entire population. The most basic set of key variables is location, sex and age. A more extensive set is location, sex, age, race and ethnicity.

28. Skinner and Shlomo (2012) use population census data from the United Kingdom to demonstrate how to estimate the risk that a record in a sample corresponds to a population unique in the census and, therefore, requires active disclosure limitation. In all disclosure limitation systems designed since Fellegi (1972) invented the discipline, records containing population uniques on key variables are the highest risk records for re-identification and receive direct disclosure avoidance protection: suppression,

coarsening categories to eliminate uniqueness, noise infusion or some combination of these. Skinner and Shlomo had to predict the probability that a sample record was a population unique because, depending on the sampling rate, records that are unique on the variables in the sample may have many duplicates in the population. They used the UK census to validate their prediction model.

29. In the case of the reconstructed 2010 Census microdata, we know the probability that a record is unique – no estimation is necessary. I presented some summary statistics on the prevalence of population uniques in the 2010 Census in my first Declaration. The location identifier is the census block code. The other two key identifiers are sex and age (in years). As I noted in my first Declaration, in the overall population, 44% of all persons are population uniques on these three variables, making them vulnerable to a classic record linkage attack identical to the one modeled by Duncan and Lambert and by Skinner and Shlomo resulting in a re-identification, when the attacker knows the name of the person associated with the location, sex and age. This is exactly the definition of a re-identification used in the McClure and Reiter paper cited by Dr. Ruggles and in the Wasserman and Zhou paper cited above. This risk assessment is derived from conventional statistical disclosure limitation methods, not differential privacy accounting.

30. Table 1 elaborates on the analysis from my first Declaration. It is based on the actual 2010 Census, not simulated data like those Dr. Ruggles uses. It uses the exact distribution of block populations found in the official Census data and the actual responses on the 2010 Census. Table 1 shows the distribution of the population by the size of the block where the person resides. Only 2.61% of the population lives in blocks with 1 to 9 persons. This is significant because these very small blocks are the ones most likely to be protected by the 2010 Census swapping method. 21.89% of the population live in blocks with 10 to 49 residents, and 22.37% live in blocks with 50 to 99 persons. Fully 46.88% of the population lives in a block with fewer than 100 residents. The column

labeled “Percent of (block, sex, age) Uniques in Bin” shows the percentage of the residents of the block who are unique in their census block, sex and age (in years) values. This percentage ranges from almost everyone (95.06%) in the least populous blocks to very few (1.12%) in the most populous blocks. There are no simulated or reconstructed data used in this table. These are characteristics of the 2010 Census resident population as they appear in the 2010 Census Edited File (CEF).<sup>9</sup>

31. The existence of documented population uniques, even one – not to mention 135 million – triggers mandatory active disclosure limitation, as documented in McKenna (2019b). If presented with a proposed public-use microdata file containing the variables: census block, sex, age (in years), race (OMB-designated coding), and ethnicity (OMB-designated coding) in 1990, 2000, 2010, or 2020, the Census Bureau Disclosure Review Board (or its predecessor) would have insisted on aggregation of the census block codes into more populous geographic areas and would have imposed minimum population sizes (at least 100,000) and minimum population thresholds for the race and ethnicity coding. It would also have insisted on sampling, as documented in McKenna (2019a).

---

<sup>9</sup> In the swapped version of the 2010 CEF, called the Hundred-percent Detail File, which was actually used for the Summary File 1 tabulations, 43.95% of the persons are population uniques using block, sex and age, almost identical to the 43.87% rate in the CEF.

Block Population Bin	Number of Blocks in Bin	2010 Census Population in Bin	Cumulative Population	Percent of Population in Bin	Cumulative Percent of Population	Population Uniques (block, sex, age) in Bin	Percent of (block, sex, age) Uniques in Bin
TOTAL	11,078,297	308,745,538				135,432,888	43.87%
0	4,871,270	0	0	0.00%	0.00%		
1-9	1,823,665	8,069,681	8,069,681	2.61%	2.61%	7,670,927	95.06%
10-49	2,671,753	67,597,683	75,667,364	21.89%	24.51%	53,435,603	79.05%
50-99	994,513	69,073,496	144,740,860	22.37%	46.88%	40,561,372	58.72%
100-249	540,455	80,020,916	224,761,776	25.92%	72.80%	27,258,556	34.06%
250-499	126,344	42,911,477	267,673,253	13.90%	86.70%	5,297,867	12.35%
500-999	40,492	27,028,992	294,702,245	8.75%	95.45%	1,051,924	3.89%
1000+	9,805	14,043,293	308,745,538	4.55%	100.00%	156,639	1.12%

DRB clearance number CBDRB-FY21-DSEP-003.

32. This table shows that whether the reconstructed 2010 Census microdata are extremely accurate, as the Census Bureau has documented, or whether “[a] much-vaunted database reconstruction technique does not perform significantly better than a crude random number generator combined with a simple assignment rule for race and ethnicity,” as Dr. Ruggles (p. 8) claims, asks the wrong question. The reconstructed data are subject to Census Bureau Disclosure Review Board regulation because they contain known population unique identifiers (the combination of census block, sex and age in years). They were produced using tabulations from a confidential Census Bureau data file – the swapped version of the CEF. And they are in record-level format with one record for every person enumerated in the 2010 Census. In their present form, they would not have been certified for release in 2011, when the other 2010 Census data products were released, nor were they certified for release in 2019, when the Census Bureau performed the full reconstruction – even though any person anywhere in the world can perform the same reconstruction because the tables *were* approved for release. The reconstructed 2010 Census data present a clear and present disclosure

risk based on the in-place standards of the Census Bureau, which predate differential privacy by several decades. They also present a clear and present disclosure risk using the traditional methods of assessing such risks, as initiated by Duncan and Lambert, refined by Skinner and Shlomo, and analyzed by the methods used in McClure and Reiter. Indeed, Dr. Ruggles' own institute, IPUMS, acknowledges that national statistical offices, like the U.S. Census Bureau, supply the microdata samples and apply disclosure limitation procedures to those data including, for recent data, limitation of the geographic detail in such microdata files even when they are samples rather than the universe.

33. The traditional standard for applying disclosure limitation methods to microdata is based on the *existence of known unique identifier combinations* in the tabulation variables—census block, sex and age in years, in this case—*not their efficacy in abetting re-identification*. Statistical agencies are expected to document the uniqueness of the identifier—that is done in my previous Declaration and in Table 1—and to continually assess the adequacy of the proposed disclosure limitation methods. Such assessments often involve re-identification studies. Such studies inform the strength of the traditional disclosure limitations applied.
34. Dr. Ruggles claims such studies are not useful because “[i]t would be impossible to positively identify the characteristics of any particular individual using the database reconstruction without access to non-public internal census information” (p. 9). The statement is false because an external agent can also conduct fieldwork or reference multiple commercially available data sources. But even more fundamentally, Dr. Ruggles' statement is irrelevant because it is the agency's duty to protect the confidentiality of the microdata and therefore it must, just as in cybersecurity, assume that attackers are clever enough to gather information that confirms the efficacy of their attacks.



35. Table 2 shows that the reconstruction-abetted re-identification attack simulated by the Census Bureau has very high precision precisely in the blocks that are most vulnerable to such an attack, whether one uses the best-case or worst-case analysis. In blocks with populations between 1 and 9 persons, the re-identification attack has a precision of 72.24% when using commercial data available in 2010.<sup>10</sup> Almost all of Dr. Ruggles' precision comes from the most populous blocks, whereas his precision plummets in sparsely populated blocks. In these sparsely populated blocks, the re-identification attack is much more precise than Ruggles' model. The exact block population was public information following the release of the 2010 Census data (as it may be in 2020 if plaintiffs succeed here). That means an attacker has a clean, public predictor of the success of the re-identification attack. Fieldwork in sparsely populated blocks can confirm this precision, as can sophisticated Bayesian methods like entity resolution without field work (Steorts, Hall and Fienberg 2016). If the attacker has better quality name, address, sex and age data than were available in 2010, certainly a plausible assumption, then the worst-case analysis for blocks with populations of 1 to 9 is precision of 96.98%--more precise than the 95% confidence interval test often used in statistics. Again, this can be confirmed by fieldwork or Bayesian entity resolution. The situation is only a little better for the 68 million people who live in blocks with populations of 10 to 49. The precision of the 2010-era commercial data is 53.61%--correct more than half the time, and the precision with high-quality external data is 91.68%. Although the best-case precisions for block populations of 50 or more are less than one-half, the worst-case precision, even in the most populous blocks, is always greater than one-half -- *an attacker with high quality external data is always more likely to be correct than wrong*. As I reported in my first Declaration, with high-quality data, the attacker

---

<sup>10</sup> Precision is the rate at which putative re-identifications are confirmed. A precision of zero indicates the putative re-identification is never correct. A precision of 100% indicates that it is always correct.

is correct on average three times out of four regardless of the number of persons who live in the block.

Block Population Bin	Putative Re-identifications (Source: Commercial Data)	Confirmed Re-identifications (Source: Commercial Data)	Precision (Source: Commercial Data)	Putative Re-identifications (Source: CEF)	Confirmed Re-identifications (Source: CEF)	Precision (Source: CEF)
TOTAL	137,709,807	52,038,366	37.79%	238,175,305	178,958,726	75.14%
0						
1-9	1,921,418	1,387,962	72.24%	4,220,571	4,093,151	96.98%
10-49	25,148,298	13,481,700	53.61%	47,352,910	43,415,168	91.68%
50-99	30,567,157	12,781,790	41.82%	51,846,547	42,515,756	82.00%
100-249	38,306,957	13,225,998	34.53%	63,258,561	45,807,270	72.41%
250-499	21,789,931	6,408,814	29.41%	35,454,412	22,902,054	64.60%
500-999	13,803,283	3,460,118	25.07%	23,280,718	13,514,134	58.05%
1000+	6,172,763	1,291,984	20.93%	12,761,586	6,711,193	52.59%
DRB clearance number CBDRB-FY21-DSEP-003.						

36. The Data Stewardship Executive Policy Committee (DSEP) determined that the simulated attack success rates in Table 2 were unacceptable for the 2020 Census. Decennial census data protected by the 2010 disclosure avoidance software is no longer safe to release. Dr. Ruggles takes issue with this conclusion for the following three reasons.

37. First, he asserts “[t]he reconstructed data are usually incorrect” (p. 10). That is not the standard. The reconstructed data are always correct for block and voting age and have at most one error – not in block or voting age – for 78% of the population. The disclosure avoidance problem is that the reconstructed data contain a known unique key (census block, sex and age in years), and therefore would be subject to the microdata disclosure avoidance rules – which mandate coarsening of the geographic identifier to areas with populations of at least 100,000.

38. Second, he asserts “[t]he reconstructed data usually do not match even the block, age and sex of anyone identified in outside commercial sources” (p. 10). That assertion

uses the wrong standard as well. What matters is the precision of the re-identification, not the absolute rate. That precision is predictably very large precisely for the population the swapping system was supposed to protect – those in sparsely populated blocks like those with a population of 1-9 people (72.24% confirmed in Commercial Data) and 10-49 people (53.61% confirmed in Commercial Data).

39. Finally, he asserts “[i]n the minority of cases where a hypothetical reconstructed individual does match the block, age and sex of someone in the commercial data, it usually turns out that the person identified in the commercial data was not actually enumerated on that block in the census” (p.10). The most favorable interpretation of his assertion is that it is based on the average precision in the best case (38%), but even under the best case, the precision is greater than half (the attacker is usually right) for the 76 million people who live on blocks with populations less than 50 people. But the Census Bureau does not calibrate its disclosure avoidance systems based on the best case because that would be irresponsible. Instead, the Census Bureau has historically relied on conservative analyses, closer to worst-case than best case, to calibrate disclosure avoidance for public-use microdata files (McKenna 2019a).
40. By relying on a simplistic and flawed analysis, Dr. Ruggles and the plaintiffs claim that reconstruction-abetted re-identification is impossible. That is wrong and accepting that view would put the privacy of millions of Americans at risk. As I showed in my prior Declaration and supplemented in this Declaration the risk to the American public from these types of attacks will certainly grow over the coming years. The risk was confirmed by non-political, career experts who serve on the Census Bureau’s Data Stewardship Executive Policy Committee. And DSEP decided – based on data – that using state-of-the-art differential privacy to implement the TopDown Algorithm was the best way to protect against those real-world threats.

I declare under penalty of perjury that the foregoing is true and correct.

DATED and SIGNED:

**JOHN ABOWD**

Digitally signed by JOHN  
ABOWD

Date: 2021.04.26 15:47:57 -04'00'

---

John M. Abowd

Chief Scientist and Associate Director for Research and Methodology

United States Bureau of the Census

## References

Cohen, A., and K. Nissim. 2020. "Linear Program Reconstruction in Practice." *Journal of Privacy and Confidentiality* 10 (1). <https://doi.org/10.29012/jpc.711>. Conference version. 2018. *Theory and Practice of Differential Privacy* [1810.05692] [Linear Program Reconstruction in Practice \(arxiv.org\)](#).

Cox, L. H. 1976. *Statistical Disclosure in Publication Hierarchies*. Report No. 14 of the research project Confidentiality in Surveys. Department of Statistics, University of Stockholm, Stockholm.

Duncan, G., and D. Lambert. 1989. "The Risk of Disclosure for Microdata." *Journal of Business and Economic Statistics*, 7(2):207-217. doi:10.2307/1391438

Evans, T., L. Zayatz, J. Slanta. 1998. "Using Noise for Disclosure Limitation of Establishment Tabular Data." *Journal of Official Statistics*, 14(4): 537. 551 [Using Noise for Disclosure Limitation of Establishment Tabular Data \(scb.se\)](#).

Fellegi, I. P. 1972. "On the question of statistical confidentiality," *Journal of the American Statistical Association*, 67:7-18.

McClure, D. and J Reiter. 2012. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." *Transactions on Data Privacy*, 5:535-552.

McKenna, L. 2019. "[Disclosure Avoidance Techniques Used for the 1960 Through 2010 Census](#)." <https://www.census.gov/library/working-papers/2019/adrm/six-decennial-censuses-da.html>. Retrieved April 23, 2021.

McKenna, L. 2019b. "U.S. Census Bureau Reidentification Studies," <https://www.census.gov/library/working-papers/2019/adrm/2019-04-ReidentificationStudies.html>. Retrieved April 23, 2021.

Skinner, C. and N. Shlomo. 2008. "Assessing Identification Risk in Survey Micro-data Using Log-Linear Models. *Journal of the American Statistical Association*," 103(483): 989-1001. Retrieved April 23, 2021, from <http://www.jstor.org/stable/27640138>

Steorts, R. C., R. Hall and S. E. Fienberg. 2016. "A Bayesian Approach to Graphical Record Linkage and Deduplication," *Journal of the American Statistical Association*, 111(516):1660-1672, DOI: 10.1080/01621459.2015.1105807.

Warner, S. L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association*, 60(309): 63-69.

Wasserman, L., and S. Zhou. 2010. "A Statistical Framework for Differential Privacy." *Journal of the American Statistical Association*, 105(489): 375-389.

# The modernization of statistical disclosure limitation at the U.S. Census Bureau

August 2020 (supersedes the 2017 version)

John M. Abowd<sup>1</sup>, Gary L. Benedetto<sup>2</sup>, Simson L. Garfinkel<sup>3</sup>, Scot A. Dahl<sup>4</sup>, Aref N. Dajani<sup>2</sup>, Matthew Graham<sup>5</sup>, Michael B. Hawes<sup>2</sup>, Vishesh Karwa<sup>6</sup>, Daniel Kifer<sup>7</sup>, Hang Kim<sup>8</sup>, Philip Leclerc<sup>2</sup>, Ashwin Machanavajjhala<sup>9</sup>, Jerome P. Reiter<sup>10</sup>, Rolando Rodriguez<sup>2</sup>, Ian M. Schmutte<sup>11</sup>, William N. Sexton<sup>12</sup>, Phyllis E. Singer<sup>2</sup>, and Lars Vilhuber<sup>2,12</sup>

<sup>1</sup> Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau, [John.Maron.Abowd@census.gov](mailto:John.Maron.Abowd@census.gov)

<sup>2</sup> Center for Enterprise Dissemination, Disclosure Avoidance, U.S. Census Bureau, [firstname.m.lastname@census.gov](mailto:firstname.m.lastname@census.gov)

<sup>3</sup> Senior Computer Scientist for Confidentiality and Data Access U.S. Census Bureau, [Simson.L.Garfinkel@census.gov](mailto:Simson.L.Garfinkel@census.gov)

<sup>4</sup> Economic Statistical Methods Division, U.S. Census Bureau, [Scot.Alan.Dahl@census.gov](mailto:Scot.Alan.Dahl@census.gov)

<sup>5</sup> Center for Economic Studies, U.S. Census Bureau, [firstname.m.lastname@census.gov](mailto:firstname.m.lastname@census.gov)

<sup>6</sup> Department of Statistics, Harvard University, [vkarwa@seas.harvard.edu](mailto:vkarwa@seas.harvard.edu)

<sup>7</sup> Department of Computer Science and Engineering, Penn State University, [dkifer@cse.psu.edu](mailto:dkifer@cse.psu.edu)

<sup>8</sup> Department of Mathematical Sciences, University of Cincinnati, [hang.kim@uc.edu](mailto:hang.kim@uc.edu)

<sup>9</sup> Department of Computer Science, Duke University, [ashwin@cs.duke.edu](mailto:ashwin@cs.duke.edu)

<sup>10</sup> Department of Statistical Science, Duke University, [jerry@stat.duke.edu](mailto:jerry@stat.duke.edu)

<sup>11</sup> Department of Economics, University of Georgia, [schmutte@uga.edu](mailto:schmutte@uga.edu)

<sup>12</sup> Labor Dynamics Institute, Cornell University, {wms32,lv39}@cornell.edu

**Abstract:** Until recently, most U.S. Census Bureau data products used traditional statistical disclosure limitation (SDL) methods such as cell or item suppression, data swapping, input noise injection, and censoring to protect respondents' confidentiality. In response to developments in mathematics and computer science since 2003 that have significantly increased the risk of reconstruction and re-identification attacks, the Census Bureau is developing formally private SDL methods to protect its data products. These methods provide mathematically provable protection for respondent data and allow policy makers to manage the tradeoff between data accuracy and privacy protection—something previously done by technical staff. The first Census Bureau product to use formal methods for privacy protection was OnTheMap, a web-based mapping and reporting application that shows where workers are employed and where they live. Recent research for OnTheMap is implementing formal privacy guarantees for businesses to complement the existing formal protections for individuals. Research is underway to improve the disclosure limitation methods for the 2020 Census of Population and Housing, the American Community Survey, and the 2022 Economic Census. For each of these programs, we are developing new state-of-the-art privacy protection approaches based on formal mechanisms that have been vetted by the scientific community. There are many challenges in adopting formally private algorithms to datasets with high dimensionality and the attendant sparsity. In addition to formally private methods that allow senior executives to set the privacy-loss budget, our implementations will feature adjustable “sliders” for allocating the privacy-loss budget among related statistical products. The Census Bureau is implementing the settings for the privacy-loss budget and these sliders based on the decisions of the Census Bureau's Data Stewardship Executive Policy Committee.

# **1 Overview: Disclosure Limitation at the U.S. Census Bureau Today**

The U.S. Census Bureau views disclosure limitation not just as a research interest, but as an operational imperative. The Census Bureau's hundreds of surveys and censuses of households, people, businesses, and establishments yield high quality data and derived statistics only if the Census Bureau maintains effective data stewardship and public trust.

The Census Bureau previously used traditional statistical disclosure limitation (SDL) techniques such as top- and bottom-coding, suppression, rounding, binning, noise injection, and sampling to preserve the confidentiality of respondent data. The Census Bureau is currently transitioning from these methods to modern SDL techniques based on formally private data publication mechanisms.

## **1.1 Legal Requirements**

The Census Bureau collects confidential information from U.S. persons and businesses under the authority of Title 13 of the U.S. Code. Once collected, the confidentiality of that data is protected specifically by 13 USC §9, which prohibits:

- (i) Using the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or
- (ii) Making any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or
- (iii) Permitting anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual records.

The privacy protections required by Title 13 are determined by the Census Bureau. Data users, including the Department of Justice and other government agencies, may be consulted regarding the criteria that determine fitness for use. Such consultation always respects the statistical-use-only requirement in the statute.

Some publications are further protected by Title 26 of the U.S. Code, which protects the federal tax information (FTI) used by the Census Bureau in the preparation of statistical products.

Confidentiality protection is intimately related to the statutory requirement that the published data be used for statistical purposes only. The definitions of "statistical purpose" and "nonstatistical purpose" were strengthened in Title III of the Foundations for Evidence-Based Policymaking Act of 2018, which is known as the Confidential Information Protection and Statistical Efficiency Act of 2018 (CIPSEA).

Additionally, the Department of Commerce (2017), in which the Census Bureau is housed, has issued directives regarding the protection of personally identifiable information (PII) and business identifiable information (BII). These directives largely mirror those issued by other government agencies and prohibit release of information

that can be used “to distinguish or trace an individual’s identity, such as their name, social security number, biometric records, etc., alone or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother’s maiden name, etc.”

### **1.2 Legacy methods supporting statistical disclosure limitation (SDL)**

Historically, the Census Bureau has primarily used information reduction and data perturbation methods to support SDL (Lauger et al., 2014). Information reduction methods include top- and bottom-coding, suppression, rounding or binning, and sampling collected units for release in public use microdata files. Data perturbation methods include swapping, legacy noise injection systems, and partially and fully synthetic database construction. These legacy approaches start with the premise that there are specific data elements that must be protected (e.g., a person’s income). A technical analyst chooses an approach from the assortment of available SDL methods that is likely to protect the data without resulting in too much damage to the published data accuracy. Usually, the selection of SDL method takes into consideration the intended uses of the published data along with assumptions about the kind of external data an intruder might have, and the types of privacy attacks an intruder might attempt.

These *ad hoc* approaches do not offer formal guarantees of data confidentiality. That is, there is no mechanism for quantifying how much privacy is being leaked from all publications based on a particular confidential database, or how one publication might interact with another publication or external data to create additional privacy risk. Furthermore, as the parameters of these legacy methods and their impact on the resulting accuracy of the data often needed to be kept confidential, there was limited opportunity for scientific scrutiny of their implementation or their effects.

### **1.3 Formal privacy approaches**

Formal privacy methods take a different approach to protecting confidential information. Instead of starting with a list of confidential values to protect, an ad hoc collection of protection mechanisms, and ad hoc assumptions about attack models, the formal approach starts with a mathematical definition and framework for quantifying privacy risk, which permits the formulation of mathematically provable privacy guarantees against unwanted inference. Next, it implements mechanisms for publishing mathematical functions (typically called *queries*) based on the confidential data that are provably consistent with the formal privacy definition. Thus, data tables released by the statistical agency are actually modeled as a series of queries applied to the confidential data. Surrogates for public use microdata files can also be generated in this manner: instead of sampling the actual respondent data, queries are used to create formally private synthetic data. This is commonly done by first modeling the confidential data, then using the model to generate synthetic data, as discussed below.

Differential privacy (Dwork et al., 2006) is the most developed formal privacy method. It begins by specifying the structure of the confidential database to be protected,  $D$ . In



computer science, this is called the database schema; in statistics, it is referred to as the sample space. Two databases,  $D_1$  and  $D_2$ , with the same schema are adjacent if the appropriately defined distance between them is, at most, unity. Leaving the technical details aside, say  $|D_1 - D_2| \leq 1$ . The universe of tables to be published from  $D$  is modeled as a set of queries on  $D$ , say  $Q$ . An element of  $Q$ , say  $q$ , is a single query on  $D$ . A randomized algorithm,  $A$ , takes as inputs  $D$ ,  $q$ , and an independent random variable. The output of  $A(D, q)$  is the statistic to be published, say  $S$ , which is a measurable set in the probability space defined by the independent random variable, say  $B$ . A randomized algorithm  $A$  for a publication system for releasing all of the queries in  $Q$  is  $\epsilon$ -differentially private if, for all  $D_1$  and  $D_2$ , with the same database schema and  $|D_1 - D_2| \leq 1$ , for all  $q \in Q$ , and for all  $S \in B$ :

$$\Pr[A(D_1, q) \in S] \leq e^\epsilon \Pr[A(D_2, q) \in S].$$

The probability is defined by the independent random variable that is used by the algorithm  $A$ , and not by the probability of observing any database  $D$  with the allowable schema (likelihood function in statistics).

There are alternative ways to define adjacent databases. For example, one method considers the databases adjacent if the record of a single person is added or removed from the database. Alternatively, the value of a single data item on a single record can be changed. Differential privacy is the mathematical formalization of the intuition that a person's privacy is protected if the statistical agency produces its outputs in a manner insensitive to the presence or absence of that person's data in the confidential database.

In differential privacy, the value  $\epsilon$  is the measure of privacy loss or confidentiality protection. If  $\epsilon = 0$ , then the two probability distributions in the definition always produce exactly the same answer from adjacent inputs—there is no difference in the output of algorithm  $A$  when given adjacent database inputs. Since the definition applies to the universe of potential inputs, and all databases adjacent to those inputs, all databases therefore produce exactly the same answer. Thus, the value  $\epsilon = 0$  guarantees no privacy loss at all (perfect confidentiality protection), but no data accuracy, since it is equivalent to releasing no data at all about the statistic  $S$ . In contrast, when  $\epsilon = \infty$ , there is no confidentiality protection at all—full loss of privacy, but the statistic  $S$  is perfectly accurate (identical to what would be produced directly from the confidential input database). Thus,  $\epsilon$  can be thought of as the *privacy-loss budget* for the publication of the queries in  $Q$ : the amount of privacy that individuals must give up in exchange for the accuracy that can be allowed in the statistical release.

Varying the privacy-loss budget allows us to move along a privacy-accuracy *Production Possibilities Frontier* (PPF) curve, as it is known in the economics literature, or along the *Receiver Operating Characteristics* (ROC) curve, as it is known in the statistics literature (Abowd and Schmutte 2019). For any attacker model, the curve constrains the aggregate disclosure risk that any confidential data might be jeopardized through any feasible reconstruction attack, given all published statistics. This budget is the worst-case limit to the inferential disclosure of any identity or item. In differential privacy,

that worst case is over all possible databases with the same schema for all individuals and items and over all external linking databases with any subset of that schema or those items.

The privacy-loss budget applies to the combination of *all* released statistics that are based on the confidential database. As a result, the formal privacy technique provides protection into the indefinite future and is not conditioned upon additional data that the attacker may have.

It is important to understand that the formal privacy protection offered by differential privacy is not absolute. Instead, it is a promise to individuals regarding the maximum amount of additional privacy loss that they may suffer as a result of a publication that is based in part on their confidential data.

To prove that a privacy-loss budget is respected, one must quantify the privacy-loss expenditure of each algorithm used to query the confidential data. The collection of the algorithms considered altogether must satisfy the privacy-loss budget. This means that the collection of algorithms used must have known composition properties.

Because the information environment is changing much faster today than when traditional SDL techniques were developed, it may no longer be reasonable to assert that a product is empirically safe given best-practice disclosure limitation prior to its release. Formal privacy models replace empirical disclosure risk assessment with designed protection. Resistance to all future attacks is a property of the design.

Differential privacy, the leading formal privacy method, is robust to background knowledge of the data, allows for sequential and parallel composability and for arbitrary post-processing edits, and enables full transparency of the implementation's source code. Differential privacy's proven guarantees hold even if external data sources are published or released later. Other formal privacy methods quantify the privacy loss that can also be mathematically established and proven, but with more constrained properties (e.g., Haney et al., 2017).

## **2 Expanding privacy protection for OnTheMap**

Randomized response, a survey technique invented in the 1960s, was the first differentially private mechanism implemented by any statistical agency. Of course, randomized response was not recognized as being differentially private until *after* differential privacy was invented. Randomized response is sometimes called *local differential privacy*. Unfortunately, it is difficult to adapt randomized response to modern survey collection methods (Wang et al., 2016). It is the Census Bureau's experience that randomized response has a poor tradeoff between accuracy and privacy protection compared with the trusted curator model, and formal assessments of the expected additive errors of the two approaches confirm this (Kasiviswanathan et al., 2011). Vadhan notes "We have a better understanding of the local model than [multi-curator models where each trusted curator holds a portion of the confidential dataset.]

However, it still lags quite far behind our understanding of the single-curator model, for example, when we want to answer a set Q of queries (as opposed to a single query).” (Vadhan 2017)

The first production application of a formally private disclosure limitation system by any organization was the Census Bureau’s OnTheMap (residential side only), a geographic query response system for studying residence and workplace patterns.

The Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES), the data used by OnTheMap, is a partially synthetic dataset that describes geographic patterns of jobs by their employment locations and residential locations as well as the connections between the two locations (U.S. Census Bureau, 2016). A job is counted if a worker is employed with positive earnings during the reference quarter and in the quarter prior to the reference quarter. These data and marginal summaries are tabulated by several categorical variables. The origin-destination (OD) matrix is made available by ten different “labor market segments”. The area characteristics (AC) data—summary margins by residence block and workplace block—contain additional variables including age, earnings, and industry. The blocks are defined in terms of 2010 Census blocks, defined for the 2010 Census of Population and Housing. The input database is a linked employer-employee database, and statistics on the workplaces (Quarterly Workforce Indicators: QWI) are protected using noise injection together with primary suppression (Abowd et al., 2009, 2012).

For OnTheMap and the underlying LODES data, the protection of the residential addresses is independent of the protection of workplaces. Protection of worker information is achieved using a formal privacy model (Machanavajjhala et al., 2008); work is in progress to protect workplaces using formal privacy as well (Haney et al., 2017).

### **3 SDL methods supporting the 2020 Census of Population and Housing**

The 2000 and 2010 Censuses of Population and Housing applied SDL in the form of record swapping, but this fact was not always obvious to data users. The actual swapping rate was kept confidential, as was the overall impact that swapping had on data accuracy (McKenna 2018).

The Census Bureau successfully tested the feasibility of producing differentially private tabulations of the redistricting data (PL94-171) for the 2018 End-to-End Census Test, and is currently in the final stages of algorithm development, for the full-scale implementation of differentially private protections for the 2020 Census of Population and Housing.

In October 2019 the Census Bureau re-released data from the 2010 Census using an early prototype for the 2020 Census Disclosure Avoidance System (DAS) (U.S. Census Bureau 2019). Called the 2010 Demonstration Data Products, this system was the subject of a December 2019 meeting of the Committee on National Statistics, where

attendees compared the statistical accuracy of these data products with previous data publications based on the 2010 Census. The source code used to prototype the 2010 Demonstration Data Products was released the following month. This code base included 33,853 lines of Python programs and 1263 lines of configuration files. In July 2020, the Census Bureau subsequently re-released the 2010 Census data protected using an updated version of the 2020 Census DAS, as the 2010 Demonstration Privacy-Protected Microdata File 2020-05-27 (U.S. Census Bureau 2020).

The differentially private mechanisms designed for the 2020 Census support the following products:

- **Public Law (PL) 94-171** files for redistricting;
- **Demographic Profiles and Demographic and Housing Characteristics files** for demographic statistics pertaining to individuals and housing units;
- **Detailed tabulations on race, ethnicity, and household composition;**
- **Privacy Protected Microdata**, the actual microdata from which published data products were tabulated; and
- **Noisy Measurements**, the actual differentially private statistics used to create the consistent microdata, to allow researchers outside the Census Bureau to produce independent statistical products without suffering the unavoidable accuracy loss that results from the post-processing of the differentially private statistics to convert them back into microdata for tabulation.

The Census Bureau has designed its differentially private algorithms to allow a selected number of queries based on the confidential data to be reported exactly. Such queries are called *invariants*. The Census Bureau currently plans the following invariants for the 2020 Census data publications:

- Total number of people by state;
- Total number of housing units (aggregate of occupied and vacant housing units) by block; and
- Total number of group quarters within three-digit group quarters type by block. Group quarters types are defined in Table P43 (U.S. Census Bureau 2012).<sup>1</sup>

While the inclusion of these invariants requires clarification of the formal privacy guarantees under differential privacy, they were considered necessary to permit public scrutiny of the state apportionment totals, and to permit the public-input component of the Local Update of Census Addresses (LUCA) program.

---

<sup>1</sup> Table P43, “Group Quarters Population by Sex and Age by Group Quarters Type,” is in Segment 6 of the 2010 Census SF1. It can be downloaded from [https://www2.census.gov/census\\_2010/04-Summary File 1/](https://www2.census.gov/census_2010/04-Summary File 1/).

Key disclosure limitation challenges include:

1. Ensuring consistency across tables by respecting the invariants enumerated above;
2. Producing block-level microdata for use by the Census Bureau's tabulation system to support production of traditional data products;
3. As was true of historical systems like swapping, there is difficulty detecting coding errors, particularly as they relate to verifying privacy-loss guarantees;
4. Determining how much of the privacy-loss budget should be spent per household; e.g., whether it should be proportional to household size;
5. A lack of high-quality usage data from which to infer relative importance of data products; and
6. The lack of public input data with which to develop and simulate the mechanism.

Key policy-related challenges include:

1. Communicating the global disclosure risk-data accuracy tradeoff effectively to the Data Stewardship Executive Policy Committee (DSEP) so that they can set the privacy-loss budget and the relative accuracy of different publications,
2. Providing effective summaries of the social benefits of privacy vs. data accuracy, so that DSEP, in particular, can understand how the public views these choices.

Throughout each decade, the Census Bureau also conducts special tabulations of small geographic areas such as towns. Those tabulations also impact privacy, and they also undergo SDL.

#### **4 SDL methods supporting the American Community Survey (ACS)**

The American Community Survey (ACS) is the successor to the long form survey of the Census of Population and Housing. The housing unit survey includes housing, household, and person-level demographic questions about a broad range of topics. There is a separate questionnaire for those residing in group quarters. The Census Bureau sends this survey to approximately 3.5 million housing units and group quarters each year and receives approximately 2.5 million responses. Weighted adjustments account for nonresponse, in-person interview subsampling, and controlling to pre-specified population totals. The ACS sample is usually selected at the tract level and is designed to allow reliable inferences for small geographic areas and for subpopulations, when cumulated across five years. ACS sampling rates vary across tracts. On average, a tract will have approximately thirty-five housing units and ninety people in the returned sample.

The Census Bureau releases one-year and five-year ACS data products. Five-year tables are released either by block group or by tract. One-year tables have been released only

for geographies containing at least 65,000 people. A recent Census Bureau Disclosure Review Board (DRB) decision allowed some one-year tables to be released for areas of at least 20,000, due to the termination of the three-year data products. The Census Bureau also releases one-year and five-year Public-Use Microdata Sample (PUMS) files for both persons and housing units. These PUMS contain samples of ACS microdata records (1% and 5% samples, respectively) with geographic detail limited to Public Use Microdata Areas (PUMA). PUMAs are special non-overlapping areas that partition each state into contiguous geographic units containing roughly 100,000 people.

The feasibility of developing formally private protection mechanisms given current methodological and computational constraints, the large number of ACS variables, and the desire for small area estimates is undemonstrated. The Census Bureau is actively pursuing this research, seeking to leverage advances from other data products. The Census Bureau is also funding cooperative agreement opportunities for research into the use of formal privacy for surveys in general. As an intermediate step to provide additional privacy to ACS respondents, the Census Bureau is experimenting with the development of non-formally private synthetic data using statistical and machine learning models to replace the current SDL methods.

Key disclosure avoidance challenges include:

1. **High dimensionality:** there are roughly two hundred topical module variables with mixed continuous and categorical values,
2. **Geography**, with estimates needed at the Census tract and block-group levels,
3. **Variable associations** across people in the same household,
4. **Outliers** in the economic variables,
5. **Survey weights** due to sampling, nonresponse, and population controls.

These challenges stem from high dimensionality combined with small sample sizes. Small geographies and sub-populations are important for data users, even if they do not always properly incorporate the sampling uncertainty when using these data. Tract-level and even block group-level data are critical for many applications, including the ballot language determinations in Section 203 of the Voting Rights Act. In addition to legislative districts, tabulations for many special geographies published by the Census Bureau, including cities and school districts, are built from smaller component geographies.

The large margins of error for small geographies allow some scope for introducing error from SDL without significantly increasing total survey error. Modelling can introduce some bias in exchange for massive decreases in variances by borrowing strength from correlations.

The research team is currently developing methods to protect ACS microdata utilizing synthesis models combined with a validation system. The overall approach is:

1. Build a chain of models, simulating each variable successively given the previous synthesized variables (Raghunathan et al., 2001). Currently, the team is assessing the use of classification trees for this purpose (Reiter, 2005);
2. Create synthetic microdata from these models for all records and all variables, creating fully synthetic data; and
3. Allow users to validate results from the synthetic microdata against the internal data. Validated results would have to meet the same standards for disclosure avoidance as all other public data releases and would be limited in quantity to statistics required for the stated purpose.

As opposed to current ACS Public Use Microdata Samples (PUMS), this fully synthetic microdata would not use internal files that have already had SDL applied to them as its source; rather, the ACS program will generate an Internal Reference File (IRF) to serve as the source. The IRF can serve as a baseline dataset for assessing survey accuracy without the confounding impacts of SDL methods, and will allow the research team to evaluate the effects of synthesis on privacy and accuracy in isolation.

The research team is considering other models for protecting tabular output, including hierarchical and spatio-temporal models.

Validation servers, verification servers<sup>2</sup>, and access to the Federal Statistical Research Data Centers (FSRDCs) may be the solution for research questions for which the modernized SDL approach leads to reasonable uncertainty regarding the suitability of published data for a particular use. An advantage of the formally private methods being tested for both the 2020 Census and the ACS is that they permit quantification of the error contributed by the SDL; hence, the inferences drawn from these data can be corrected for the impact of the uncertainty added to protect privacy. Their suitability for use in a particular application can also be assessed without reference to the confidential data. This property of modernized SDL provides a means for applying objective criteria to a researcher's claim that the published data are suitable or unsuitable for a particular use.

## **5 SDL research supporting the 2022 Economic Census**

Every five years the Census Bureau sends survey forms to nearly four million U.S. business establishments, broadly representative of all geographic regions and most private industries, to conduct the Economic Census. The Economic Census is based on a complete enumeration for certain types of businesses, and sampling of other, mostly smaller, businesses. The Census Bureau defines an *establishment* as a specific economic activity conducted at a specific location, and asks companies to file separate reports for

---

<sup>2</sup> Validation servers provide the data user with the results of their query calculated on the internal data with SDL performed on the result. Verification servers provide the data user with some measure of how confident they should be with the result of their query calculated on the synthetic data.

different locations and when multiple lines of activity are present at the same location. The Economic Census survey collects information from sampled establishments on the revenue obtained from product sales in the industries in which they operate, as well as information on employment, payroll, and other establishment characteristics.

Key policy challenges include:

1. Specifying the entity to be protected: multi-unit companies operate many establishments with different forms. From a legal standpoint, it is not entirely clear which entity (company, establishment, or something else) must be protected.
2. Defining what constitutes sufficient protection. Requirements to protect fact-of-filing may imply that whether a given business appears must be protected. However, it may not be necessary to protect certain business attributes that are in the public domain.

Key disclosure avoidance challenges include:

1. **Outliers** in the economic variables and generally high skewness;
2. **Sparsity** of data in cells disaggregated down to the North American Industry Classification System (NAICS) subsector and county level;
3. **Hybrid** sampling and enumeration design combined with an edit and imputation stage that complicate privacy models;
4. **Associations** among economic variables that increase disclosure risk; and
5. **Complex** publication schedules that require consistency over time and efficient allocation of privacy-loss budgets across releases.

The Census Bureau's disclosure modernization efforts for the Economic Census have followed two potentially complementary paths. Beginning in 2017, an interdisciplinary team at the Census Bureau partnered with academic colleagues to evaluate the feasibility of developing synthetic industry-level microdata. The methods under consideration are not formally private, but would allow publication of more detailed information while maintaining disclosure protections comparable to the cell suppression methods currently in use. Kim, Reiter, and Karr (2016) present methods of developing synthetic data on historic Economic Census data from the manufacturing sector. An inter-divisional team has applied two synthetic data models to 42 industries from the 2012 Economic Census covering eighteen economic sectors. Input data were limited to full-year reporter businesses (births, deaths, and seasonal businesses were excluded). The synthetic data were evaluated for fidelity in summary tabulations of items collected for all sectors. The team is currently evaluating the disclosure risk for these approaches. Kim and



Thompson are working on a separate synthetic data model that includes businesses that are part-year reporters.

In 2020 an additional team began work to develop formally private disclosure avoidance methods appropriate to economic data in general, and the Economic Census in particular. Since the publication schedule does not require release of microdata, the team is exploring modifications of the differential privacy paradigm that could be directly applied to tabular summaries and yield provable privacy guarantees. Specifically, they are considering a variant of the model developed in Haney et al., (2017) as well as other approaches in the smooth sensitivity framework (e.g. Nissim, Raskhodnikova and Smith, 2007). The sparsity of the published tables may require a modification of these methods to ensure consistency and data quality while keeping privacy loss at acceptable levels. The team intends to develop methods applicable to the County Business Patterns and Economic Census First Look products, which have relatively simple structure. From there it will hopefully be possible to adapt those methods to more complex Economic Census products.

## **6 Challenges and meetings those challenges**

In differential privacy, the commonly used flattened histogram representation of the universe is calculated as the Cartesian product of all potential combinations of responses for all variables. This representation is often orders of magnitude larger than the total population even when structural zeroes (impossible combinations of values of variables, such as grandmothers who are three years of age) are imposed. One promising approach is approximate differential privacy, where the limiting factor depends only on the logarithm of the inverse probability of algorithmic failure.

Policy makers, including the Census Bureau's DSEP, must have enough information about the privacy-loss/data accuracy trade-off to make an informed decision about  $\epsilon$ , and its allocation to different tabular summaries. In some cases, the chosen amount of noise injection from differential privacy may limit the suitability for use of the published statistics to more narrowly defined domains than has historically been the case.

The strategy for producing the tabular summaries is to supply the official tabulation software with formally private synthetic data that reproduce all of the protected tabulations specified in the redistricting and summary file requirements. In generating high quality synthetic microdata, one needs to consider integer counts, non-negativity, unprotected counts (e.g., total state population), and structural zeroes.

To execute this approach, the Census Bureau needs generic methods that will work on a broader range of datasets. In addition, it may be difficult to find meaningful correlations that are not represented in the model. To address this, the model must anticipate the types of analyses that data users might wish to conduct. As a result, better model-building tools are needed, as well as generic tools for correlating arbitrary models

with the ones used to build the synthetic data. Ongoing engagement with data users is also essential to help identify these intended uses of the published data.

Reproducible-science methods will be required to use synthetic data effectively.

Data are often collected with a complex sample design with considerable missing data and in panels of longitudinal data. Research is ongoing to ensure that weighted, longitudinal analysis using differentially private data will continue to produce “good results and good science” to the data users.

## **7 Approaches to gauge data accuracy and usefulness**

There are multiple methods to assess data accuracy, also known as analytical (or inference) validity. Machanavajjhala et al. (2008) conducted experiments comparing differentially private synthetic data to the actual data for OnTheMap. They saw value in coarsening the domain to limit the number of “strange fictitious commuting patterns.” Karr et al. (2006) and Drechsler (2011) advocate calculating confidence interval overlaps for parameters of interest, whether univariate, bivariate, or multivariate.

There is value in calculating all such metrics described above for parameter estimates calculated from:

- non-perturbed data (exact counts) where we expect parity; and
- parameter estimates that were not captured in the joint distributions modeled in the synthetic data, where one would not expect to uncover comparable results.

Disclosure limitation is a technology. It shows the relationship between privacy loss, which is considered a public “bad”, and data accuracy, which is considered a public “good”. A differentially private system can publish extremely disclosive data. This happens if the privacy-loss budget is set very high. The extremely disclosive data will likely be very accurate. That is, inferences based on these data will be nearly identical to those based on the confidential data. But extremely disclosive, albeit formally private, data also permit a very accurate reconstruction of the confidential data relative to the reconstruction possible with smaller privacy-loss budgets.

The teams at the Census Bureau working on formal privacy methods for statistical disclosure limitation have been charged by DSEP with developing technologies with adjustable parameters to control the privacy loss and data accuracy during implementation. Those technologies will be summarized with a variety of supporting materials. The Disclosure Review Board will make a recommendation regarding the appropriate formal privacy technology and parameter settings, including the privacy-loss parameter  $\epsilon$ . The Data Stewardship Executive Policy Committee will review that recommendation and make the final determination. The published data will implement the recommendations of DSEP. Although more explicit than in previous censuses, this is the same chain of recommendation and approval that was used in 2000 and 2010.

This transition to innovation involves significant retooling of methods for the Census Bureau’s career mathematical statisticians, computer scientists, subject matter experts, project and process managers, and internal stakeholders. This transition will help the Census Bureau lead similar innovation across the U.S. Federal Government and beyond.

## 8 References

- Abowd, John M. and Ian M. Schmutte “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices,” *American Economic Review*, Vol. 109, No. 1 (January 2019):171-202, DOI:10.1257/aer.20170627.
- Abowd, John M., R. Kaj Gittings, Kevin L. McKinney, Bryce Stephens, Lars Vilhuber, and Simon D. Woodcock (2012). *Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures for Related Time Series*. 12-13. U.S. Census Bureau, Center for Economic Studies.
- Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon D Woodcock (2009). *The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators*. In *Producer Dynamics: New Evidence from Microdata*, edited by Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. University of Chicago Press.
- Department of Commerce, Office of Privacy and Open Government (2017). *Safeguarding Information*. [http://osec.doc.gov/opog/privacy/pii\\_bii.html#PII](http://osec.doc.gov/opog/privacy/pii_bii.html#PII)
- Drechsler, Jörg (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. New York: Springer.
- Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006) Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third conference on Theory of Cryptography (TCC'06)*, Shai Halevi and Tal Rabin (Eds.). Springer-Verlag, Berlin, Heidelberg, 265-284. DOI=[http://dx.doi.org/10.1007/11681878\\_14](http://dx.doi.org/10.1007/11681878_14)
- Garfinkel, Simson, John M. Abowd, and Christian Martindale, Understanding Database Reconstruction Attacks on Public Data, *Communications of the ACM*, February 2019.
- Garfinkel, Simson, John M. Abowd, Sarah Powazek, Issues Encountered Deploying Differential Privacy, Workshop on Privacy in the Electronic Society, Toronto, Canada - October 15, 2018.
- Haney, Samuel, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber (2017). *Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics*, SIGMOD’17, May 14-19, 2017, Chicago, Illinois, USA, DOI: 10.1145/3035918.3035940.

- Karr, A.F., C.N. Kohnen, A. Oganian, J.P. Reiter, and A.P. Sanil (2006). *A framework for evaluating the utility of data altered to protect confidentiality*. *The American Statistician* 60, 224-232.
- Kasiviswanathan, Shiva Prasad, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith (2011). *What can we learn privately?*. *SIAM Journal on Computing* 40, no. 3: 793-826.
- Kim, Hang J., Jerome P. Reiter, and Alan F. Karr (2016). *Simultaneously Edit-Imputation and Disclosure Limitation for Business Establishment Data*. *Journal of Applied Statistics* online: 1-20.
- Lauger, Amy, Billy Wisniewski, and Laura McKenna (2014). *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research*. Research Report Series (Disclosure Avoidance #2014-02). Washington: Center for Disclosure Avoidance Research, U.S. Census Bureau.
- McKenna, Laura (2018). *Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing*. Working Papers 18-47, Washington: Center for Economic Studies, U.S. Census Bureau.
- Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber (2008). *Privacy: Theory Meets Practice on the Map*. Proceedings: International Conference on Data Engineering. Washington, DC, USA: IEEE Computer Society, 277-286.
- Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk, and Peter Solenberger (2001). *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models*. *Survey Methodology* 27(1). Citeseer: 85-96.
- U.S. Census Bureau (2012). 2010 Census Summary File 1: 2010 Census of Population of Housing. September 2012. U.S. Census Bureau. <https://www.census.gov/prod/cen2010/doc/sf1.pdf>
- U.S. Census Bureau (2016). OnTheMap: Data Overview (LODES Version 7). U.S. Census Bureau. <https://lehd.ces.census.gov/doc/help/onthemap/OnTheMapDataOverview.pdf>
- [U.S. Census Bureau \(2019\). 2010 Demonstration Data Product. October 2019. U.S. Census Bureau. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html)
- [U.S. Census Bureau \(2020\). 2010 Demonstration Privacy-Protected Microdata File 2020-05-27. July 2020. U.S. Census Bureau. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf/?#](https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf/?#)

- Vadhan, Salil (2017). *The Complexity of Differential Privacy*. March 14, 2017. [https://privacytools.seas.harvard.edu/files/privacytools/files/complexityprivacy\\_1\\_0\\_1.pdf](https://privacytools.seas.harvard.edu/files/privacytools/files/complexityprivacy_1_0_1.pdf)
- Vilhuber, Lars and Ian M. Schmutte (2016). *Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy*. <http://digitalcommons.ilr.cornell.edu/ldi/33/>
- Wang, Yue, Xintao Wu, and Donghui Hu (2016). *Using Randomized Response for Differential Privacy Preserving Data Collection*. Workshop proceedings of the EDBT/ICDT 2016 Joint Conference. March 15, 2016, Bordeaux, France. <http://ceur-ws.org/Vol-1558/paper35.pdf>

## 9 Disclaimer

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

# Census TopDown: The Impacts of Differential Privacy on Redistricting

Aloni Cohen ✉

Hariri Institute for Computing and School of Law, Boston University, USA

Moon Duchin ✉

Department of Mathematics, Tufts University, USA

JN Matthews ✉

Tisch College of Civic Life, Tufts University, USA

Bhushan Suwal ✉

Tisch College of Civic Life, Tufts University, USA

## Abstract

The 2020 Decennial Census will be released with a new disclosure avoidance system in place, putting *differential privacy* in the spotlight for a wide range of data users. We consider several key applications of Census data in redistricting, developing tools and demonstrations for practitioners who are concerned about the impacts of this new noising algorithm called **TopDown**. Based on a close look at reconstructed Texas data, we find reassuring evidence that **TopDown** will not threaten the ability to produce districts with tolerable population balance or to detect signals of racial polarization for Voting Rights Act enforcement.

**2012 ACM Subject Classification** Security and privacy; Applied computing → Law; Applied computing → Voting / election technologies

**Keywords and phrases** Census, TopDown, differential privacy, redistricting, Voting Rights Act

**Digital Object Identifier** [10.4230/LIPIcs.FORC.2021.5](https://doi.org/10.4230/LIPIcs.FORC.2021.5)

**Supplementary Material** <https://megg.org/DP>

**Funding** This project was supported on NSF OIA-1937095 (Convergence Accelerator) and by a grant from the Alfred P. Sloan Foundation.

*Aloni Cohen*: NSF CNS-1414119 and CNS-1915763; DARPA HR00112020021

*Moon Duchin*: NSF DMS-2005512

**Acknowledgements** Authors are listed alphabetically. We thank Denis Kazakov, Mark Hansen, and Peter Wayner. Kazakov developed the reconstruction algorithm as a member of Hansen’s research group. Wayner guided our deployment of **TopDown** in AWS and was an invaluable team member for the technical report. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our funders.

## 1 Introduction

A new disclosure avoidance system is coming to the Census: the 2020 Decennial Census releases will use an algorithm called **TopDown** to protect the data from increasingly feasible *reconstruction attacks* [2]. Census data is structured in a nesting sequence of geographic units covering the whole country, from nation at the top to small *census blocks* at the bottom. **TopDown** starts by setting a *privacy budget*  $\varepsilon > 0$  which is allocated to the levels of a designated hierarchy, then adding noise at each level in a *differentially private* way [12]. When  $\varepsilon \rightarrow \infty$ , the data alterations vanish, while  $\varepsilon \rightarrow 0$  yields pure noise with no fidelity to the input data. The algorithm continues with a post-processing step that leaves an output dataset that is designed to be suitable for public use.



© Aloni Cohen, Moon Duchin, JN Matthews, and Bhushan Suwal;  
licensed under Creative Commons License CC-BY 4.0  
2nd Symposium on Foundations of Responsible Computing (FORC 2021).

Editors: Katrina Ligett and Swati Gupta; Article No. 5; pp. 5:1–5:22



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 5:2 Census TopDown: The Impacts of Differential Privacy on Redistricting

43 *Redistricting* is the process of dividing a polity into territorially delimited pieces in which  
44 elections will be conducted. The Census has a special release—named the PL 94-171 after  
45 the law that requires it—that reports the number of residents in every geographic unit in  
46 the country by race, ethnicity, and the number of voting-age residents [9]. The 2020 release  
47 is slated to occur by September 2021, after which many thousands of district lines will  
48 be redrawn: not only U.S. Congressional districts, but those for state legislatures, county  
49 commissions, city councils, and many more.

50 Many user groups have expressed concerns about the effects of differential privacy on  
51 redistricting. They largely but not exclusively concern two issues. First, “One Person, One  
52 Vote” case law calls for balancing population across the electoral districts in a jurisdiction,  
53 whether small like city council districts or large like congressional districts. Most states  
54 balance congressional districts to within one person based on Census counts. Second, the  
55 most reliable legal tool against gerrymandering has been the Voting Rights Act of 1965  
56 (VRA), which requires a demonstration of racially polarized voting (RPV). This RPV analysis  
57 is typically performed by statistical techniques that infer voting by race from precinct-level  
58 returns. Many voting rights advocates worry that noising of Census data will confuse  
59 population balancing practices, and others worry that it will attenuate RPV signals, making  
60 it harder to press valid claims.

61 The Census Bureau has been commendably transparent about the development of  
62 **TopDown**, making working code publicly available along with documentation and research  
63 papers describing the algorithm. The complexity of the algorithm makes it extremely difficult  
64 to study analytically, so many people have sought to run it on realistic data. However, since  
65 person-level Census data remain confidential for 72 years after collection, detailed input data  
66 for **TopDown** is not public. Data users who would like to understand its impacts are left with  
67 two options: decades-old data or a limited demonstration data product.

68 In this paper, we get around the empirical obstacle by use of reconstructed block-level 2010  
69 microdata for the state of Texas, and we try to understand the algorithm through theoretical  
70 analysis of a much-simplified toy algorithm, **ToyDown**, that retains the two-stage, top-down  
71 structure of **TopDown** but is much easier to analyze symbolically. We investigate three  
72 questions about the count discrepancies created by **TopDown** in units of census geography  
73 and “off-spine” aggregations like districts and precincts.

74 **Hierarchical budget allocation.** We derive easy-to-evaluate expressions for **ToyDown** errors  
75 as a function of the privacy budget allocation. Error at higher levels of the geographic  
76 hierarchy impacts lower-level counts with a significant discount, suggesting that bottom-  
77 heavy allocations may be optimal for accuracy on small geographies. This is consistent with  
78 the small-district errors in our experiments with **TopDown**. For larger districts, a tract-heavy  
79 allocation gives greatest accuracy. Equal allocation over the levels is a strong performer in  
80 both cases, making this a good choice from the point of view of multi-scale redistricting.

81 **District construction.** From there, we create further tests to study the impacts of district  
82 design. We compare hierarchically greedy to geometrically greedy district-generation schemes,  
83 where the former attempt to keep large units whole and the latter attempt to build districts  
84 with short boundaries. We find that the **ToyDown** model gives errors very closely keyed to  
85 the fragmentation of the hierarchy, but that spatial factors damp out the primary role of  
86 fragmentation in the shift to the **TopDown** setting.

87 **Robustness of linear regression.** Finally, we consider the unweighted linear regressions  
88 commonly used to assess racial polarization in voting rights cases. We find that the noise  
89 from both **ToyDown** and **TopDown** introduces an attenuation bias that seems alarming at  
90 first. However, unweighted linear regression on precincts is already vulnerable to major skews

91 imposed by the inclusion of very small precincts. For any reasonable way of counteracting  
92 that—trimming out the tiny precincts or weighting the regression by the number of votes  
93 cast—the instability introduced by ToyDown and TopDown all but vanishes.

94 Our investigation is set up to answer questions about the status quo workflow in  
95 redistricting. As usual with studies of differential privacy, a finding that DP unsettles the  
96 current practices might lead us to call to refine the way it is applied, but might equally lead  
97 us to interrogate the traditional practices and seek next-generation methods for redistricting.  
98 In particular, it is clear that the practice of *one-person* population deviation across districts  
99 was never reasonably justified by the accuracy of Census data nor required by law, and the  
100 adoption of differential privacy might give redistricters occasion to reconsider that practice.  
101 We make a similar observation about the way that racially polarized voting analysis is  
102 commonly performed in expert reports. On the other hand, by focusing on decisions still to  
103 be announced like the privacy budget and its allocation over the hierarchy, we are able to  
104 make recommendations that can assist the Bureau in protecting privacy while attending to  
105 the important concerns of user groups.

## 106 **2 Background on Census and redistricting**

### 107 **2.1 The structure of Census data and the redistricting data products**

108 Every ten years the U.S. Census Bureau attempts a comprehensive collection of person-level  
109 data—called *microdata*—from every household in the country. The microdata are confidential,  
110 and are only published in aggregated tables subject to disclosure avoidance controls. The  
111 Decennial Census records information on the sex, age, race, and ethnicity for each member of  
112 each household, using categories set by the Office of Management and Budget [8]. The 2020  
113 Census used six primary racial categories: White, Black, American Indian, Asian, Native  
114 Hawaiian/Pacific Islander, and Some Other Race. An individual can select these in any  
115 combination but must choose at least one, creating  $2^6 - 1 = 63$  possible choices of race.  
116 Separately, *ethnicity* is represented as a binary choice of Hispanic/Latino or not.

117 The 2010 Census divided the nation into over 11 million small units called *census blocks*  
118 which nest in larger geographies in a six-level “central spine”: nation—state—county—  
119 tract—block group—block. Counts of different types are provided with respect to these  
120 geographies. This tabular data is then used in an enormous range of official capacities, from  
121 the apportionment of seats in the U.S. House of Representatives to the allocation of many  
122 streams of federal and state funding. The redistricting (PL 94-171) data includes four such  
123 tables: H1, a table of housing units whose types are occupied/vacant; and four tables of  
124 population, P1 (63 races), P2 (Hispanic, and 63 races of non-Hispanic population), and  
125 P3/P4 (same as P1/P2 but for voting age population). Each table can be thought of as a  
126 *histogram*, with each included type constituting one histogram *bin*. For instance, in table P1  
127 there is 1 person in the  $t = \text{White} + \text{Asian}$  bin in the Middlesex County, MA, block numbered  
128 31021002.

129 Treating the 2010 tables as accurate, it is easy to infer information not explicitly presented  
130 in the tables. For instance, the same bin in the P3 table (race for voting age population) also  
131 has a count of 1, implying that there are no White+Asian people under 18 years old in block  
132 31021002. This is the beginning of a *reconstruction* process that would enable an attacker, in  
133 principle, to learn much of the person-level microdata behind the aggregate releases.



## 134 2.2 Disclosure avoidance

135 Title 13 of the U.S. Code requires the Bureau to take measures to protect the privacy of  
 136 respondents' data [1]. In the 2010 Census, this was largely achieved by an ad hoc mechanism  
 137 called *data swapping*: a Bureau employee manually swapped data between small census  
 138 blocks to thwart re-identification. In 2020, swapping is no longer considered adequate to  
 139 protect against more sophisticated (but mathematically straightforward) data attacks that  
 140 seek to reconstruct the individual microdata. An internal Census Bureau study concluded  
 141 that data swapping was unacceptably vulnerable: Census staff were able to reconstruct the  
 142 2010 Census responses of—and correctly reidentify—tens of millions of people.

143 With the reconstruction/reidentification threat in mind, the Bureau has developed an  
 144 algorithm called TopDown [2], which begins with a noising step that is *differentially private*,  
 145 following a mathematical formalism that provides rigorous guarantees against information  
 146 disclosure [12]. Differentially private algorithms obey a quantifiable limit to how much the  
 147 output can depend on an individual record in the input. The relationship of output to input  
 148 is specified by a tuneable parameter,  $\epsilon$ , often called the *privacy budget*. When  $\epsilon \rightarrow \infty$ , the  
 149 output approaches equality to the input (high risk of disclosure). When  $\epsilon \rightarrow 0$ , the output  
 150 bears no resemblance to the input whatsoever (no risk of disclosure). Like a fiscal budget,  
 151 the privacy budget can be allocated until it is fully spent, in this case by spending parts of  
 152 the budget on particular queries and on levels of the hierarchy.

153 TopDown takes an individual-level table of census data and creates a ‘synthetic’ dataset  
 154 that will be used in its place to generate the PL 94-171 tables. It can be thought of as  
 155 taking as input a histogram with a bin for each person type (i.e., a combination of race, sex,  
 156 ethnicity, etc.) and outputting an altered version of the same histogram. It proceeds in two  
 157 stages. First, it privatizes the input histogram counts: it adds enough random noise to get  
 158 the required level of differential privacy (according to the budget  $\epsilon$ ). At this stage, it also  
 159 allocates a portion of the total privacy budget for generating additional noisy histograms of  
 160 data of particular importance to the Census Bureau. Second, TopDown does post-processing  
 161 on the noisy histograms to satisfy a handful of additional plausibility constraints. Among  
 162 other things, post-processing ensures that the resulting histograms contain only non-negative  
 163 integers, are self-consistent, and agree with the raw input data on a handful of *invariants*  
 164 (e.g., total state population).

165 The overall privacy guarantees of TopDown are poorly understood. In this paper, we  
 166 design a simpler cousin of TopDown nicknamed ToyDown and we explore the properties of  
 167 both ToyDown and TopDown, primarily focusing on reconstructed Texas data from 2010.

## 168 2.3 The use of Census products for redistricting

169 The PL 94-171 tables are the authoritative source of data for the purposes of apportionment  
 170 to the U.S. House of Representatives, and with a very small number of exceptions also for  
 171 within-state legislative apportionment. The most famous use of population counts is to  
 172 decide how many members of the 435-seat House of Representatives are assigned to each  
 173 state. In “One person, one vote” jurisprudence initiated in the *Reynolds v. Sims* case of  
 174 1964, balancing Census population is required not only for Congressional districts within  
 175 a state but also for districts that elect to a state legislature, a county commission, a city  
 176 council or school board, and so on [17, 18, 3].

177 Today, the Congressional districts within a state usually balance total population extremely  
 178 tightly: each of Alabama’s seven Congressional districts drawn after the 2010 Census has  
 179 a total population of either 682,819 or 682,820 according to official definitions of districts

180 and the Table P1 count, while Massachusetts districts all have a population of 727,514 or  
 181 727,515. Astonishingly, though no official rule demands it, more than half of the states  
 182 maintain this “zero-balancing” practice (no more than one person deviation) for Congressional  
 183 districts [16]. This ingrained habit of zero-balancing districts to protect from the possibility  
 184 of a malapportionment challenge is the first source of worry in the redistricting sphere. If  
 185 disclosure avoidance practices introduce some systematic bias—say by creating significant  
 186 net redistribution towards rural and away from urban areas—then it becomes hard to control  
 187 overall malapportionment, which could in principle trigger constitutional scrutiny. In the  
 188 end, redistricters may not care very much how many people live in a single census block, but  
 189 it could be quite important to have good accuracy at the level of a district.

190 The second major locus of concern for redistricting practitioners is the enforcement of the  
 191 Voting Rights Act (VRA). Here, histogram data is used to estimate the share of voting age  
 192 population held by members of minority racial and ethnic groups. Voting rights attorneys  
 193 must start by satisfying three threshold tests without which no suit can go forward.

- 194 ■ **Gingles 1:** the first “Gingles factor” in VRA liability is satisfied by creating a demonstration  
 195 district where the minority group makes up over 50% of the voting age population.
- 196 ■ **Gingles 2-3:** the voting patterns in the disputed area must display *racial polarization*.  
 197 The minority population is shown to be cohesive in its candidates of choice, and bloc  
 198 voting by the majority prevents these candidates from being elected. In practice, inference  
 199 techniques like linear regression or so-called “ecological inference” are used to estimate  
 200 voting preferences by race.

201 Since the VRA has been a powerful tool against gerrymandering for over 50 years, many  
 202 worry that even where the raw data would clear the Gingles preconditions, the noised data  
 203 will tend towards uniformity—blocking deserving plaintiffs from a cause of action.

## 204 **3** Census TopDown and ToyDown

### 205 **3.1** Setup and notation

206 For the Census application, the data universe is a set of *types*: for instance, the redistricting  
 207 data (the PL 94-171) has the types  $T = T_R \times T_E \times T_{VA} \times T_H$ , where  $T_R$  is the set of 63  
 208 races,  $T_E$  is binary for ethnicity (Hispanic or not),  $T_A$  is binary for age (voting age or not),  
 209 and  $T_H$  is the set of housing types. (The fuller decennial Census data has more types.)

210 A *hierarchy*  $H$  is a rooted tree of some depth  $d$ , so that every leaf has distance  $\leq d - 1$   
 211 from the root. We will usually assume the hierarchy has uniform depth, so that every leaf is  
 212 exactly  $d - 1$  away from the root. For node  $h \in H$ , let  $n(h) \in \mathbb{N}$  be the number of children  
 213 of  $h$  in the tree, and let  $\ell(h)$  be the level of node  $h$ . A hierarchy is called *homogeneous*  
 214 if each node at level  $\ell$  has the same number of children, denoted  $n_\ell$ . Let  $H_\ell$  denote the  
 215 set of nodes at level  $\ell$ , so that the set of leaves is  $H_d$  in the uniform-depth case. Label  
 216 the root of the tree  $h = 1$ . We adopt an indexing of the tree and refer to the  $i$ th child of  
 217  $h$  as  $h_i$ ; the parent of any non-root node  $h$  is denoted  $\hat{h}$ . In Census data, the hierarchy  
 218 represents the large and complicated set of nested geographical units, from the nation at  
 219 the root down to the census blocks at the leaves. The standard hierarchy has the six levels  
 220 (nation—state—county—tract—block group—block) described above.

221 We associate with hierarchy  $H$  and types  $T$  a set of *counts*  $A_{H,T} = \{a_{h,t} \in \mathbb{N}\}_{h \in H, t \in T}$ ,  
 222 where  $a_{h,t}$  is the population of type  $t$  in unit  $h$  of census geography. We say  $A_{H,T}$  is  
 223 *hierarchically consistent* if the counts add up correctly: for every non-leaf  $h$  and every  $t$ , we  
 224 require  $a_{h,t} = \sum_{i \in [n(h)]} a_{h_i,t}$ . For a singleton  $T$ , we write  $A_H = \{a_h\}$ . We set an *allocation*  
 225  $(\varepsilon_1, \dots, \varepsilon_d)$  breaking down the privacy budget  $\varepsilon = \sum \varepsilon_i$  to the different levels of the hierarchy.

Our *queries* will always be counting queries, so that for instance  $q_{F,44}(h)$  returns the number of 44-year-old females in geographic unit  $h$ . This particular query is part of a “sex by age” *histogram*  $Q_{sex,age} = \{q_{s,a} : s \in T_S, a \in T_A\}$ , which partitions  $T$  into *bins* by sex and age. In this language,  $q_{F,44}$  is a bin of the sex-by-age histogram. By slight abuse of notation, we will use the same terminology for the queries and their outputs, so that the histogram can be thought of as the collection of queries or the collection of counts. Similarly, the “voting age by ethnicity by race” histogram consists of a query for each combination of the  $2 \times 2 \times 63$  possible combinations of the three attributes.

### 3.2 ToyDown and TopDown

The Bureau’s TopDown and our simplified ToyDown are both algorithms for releasing privatized population counts for every  $h \in H$ . That is, these algorithms protect privacy by noising the data histograms. TopDown releases not just total population counts, but counts by type. We will define *single-attribute* and *multi-attribute* versions of ToyDown that noise  $A_H$  and  $A_{H,T}$ , respectively, where consistency must hold for each type  $t$ .

TopDown and ToyDown share the same two-stage structure. Starting with hierarchically consistent raw counts  $a$ , the *noising stage* generates differentially private counts  $\hat{a}$ . The *post-processing stage* solves a constrained optimization problem to find noisy counts  $\alpha$  that are close to the  $\hat{a}$  values while satisfying hierarchical consistency and other requirements. TopDown is named after the iterative approach to post-processing: one geographic level at a time, starting at the top (nation) and working down to the leaves (blocks). We sketch the noising and post-processing here, and we describe them in Appendix A in more detail.

The simple ToyDown model can be run in a single-attribute version (only counts  $A_H$ ), a multi-attribute version (counts by type  $A_{H,T}$ ), or in multi-attribute form enforcing non-negativity. The single-attribute version is easy to describe: level by level, random noise values are selected from a Laplace distribution with scale  $1/\varepsilon_\ell$  and added to each count, replacing each  $a_h$  with  $\hat{a}_h = a_h + L_h$ . Then, working from top to bottom, the noisy  $\hat{a}_h$  are replaced with the closest possible real numbers  $\alpha_h$  satisfying hierarchical consistency. Multi-attribute ToyDown is defined analogously, but using  $A_{H,T}$  instead of  $A_H$  and requiring hierarchical consistency within each type  $t \in T$ . Non-negative ToyDown adds the inequality requirement that  $\alpha_h \geq 0$ .

TopDown is structurally similar but much more complex, with more kinds of privatized counts in the noising stage and a great many more constraints in the post-processing stage, including integrality. The privatized counts computed by TopDown are specified by a collection of histograms (or complex queries) called a *workload*  $W$ . For each bin of each histogram in the workload and for each node  $h$  in the geographic hierarchy, TopDown adds geometric noise to the count. The post-processing step finds the closest integer point that satisfies the requirements given by hierarchical consistency, non-negativity, as well as additional conditions given as invariants and structural inequalities. For example, any block with zero households in the raw counts must have zero households and zero population in the output adjusted counts. Together, the invariants, structural inequalities, integrality, and non-negativity make this optimization problem very hard. The problem is NP-hard in the worst case and TopDown cannot always find a feasible solution. There is a sophisticated secondary algorithm for finding approximate solutions that is beyond the scope of this paper.

ToyDown is simple enough that solutions can often be obtained symbolically. ToyDown simplifies the noising stage by fixing the workload to be the detailed workload partition  $Q_{detailed} = \{\{t\}\}_{t \in T}$  consisting of all singleton sets and using the continuous Laplace Mechanism instead of the discrete Geometric Mechanism. It simplifies the post-processing

273 stage by dropping invariants, structural inequalities, integrality, and non-negativity. When  
274 negative answers are permitted, multi-attribute `ToyDown` is equivalent to executing  $|T|$   
275 independent instances of single-attribute `ToyDown` on inputs  $A_{H,t} = \{a_{h,t}\}_{h \in H}$  for each  
276  $t \in T$ . As a result, many of our analytical results for single-attribute `ToyDown` extend  
277 straightforwardly to multi-attribute `ToyDown` (allowing negative answers) by scaling by a  
278 factor of  $|T|$  in appropriate places.

## 279 4 Methods

280 We use both analytical and empirical techniques in this work. This section describes our  
281 high-level empirical approach: what algorithms and raw data we used and how we used  
282 them. See Appendix B for more details. We repeatedly ran `TopDown` and `ToyDown` in  
283 various configurations on a reconstructed person-level Texas dataset created by applying a  
284 reconstruction technique to the block-level data from the 2010 Census, following [15] based on  
285 [11]. The reconstructed microdata records—obtained from collaborators—contain block-level  
286 sex, age, ethnicity, and race information consistent with a collection of tables from 2010  
287 Census Summary File 1.

288 We executed 16 runs of `TopDown` with each of 20 different allocations of the privacy budget  
289 across the five lower levels of the national census geographic hierarchy:  $\varepsilon = \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_6$ .  
290 The 20 allocations consist of five different splits across the levels (Table 1) for each of four  
291 total budgets  $\varepsilon \in \{0.25, 0.5, 1.0, 2.0\}$ . `TopDown` operates on the six-level Census hierarchy  
292 and requires specifying  $\varepsilon_1$ . In our experiments, we ran `TopDown` with a fixed total privacy  
293 budget  $\varepsilon_{total} = 10$ , with  $\varepsilon_1 = 10 - \varepsilon$ . Because the nation-level budget is so much higher  
294 than the lower level budgets, we omit further discussion of it. The `TopDown` workload was  
295 modeled after the workload used in the 2018 End-to-End test release, omitting household  
296 invariants and queries.

297 We also ran three variants of `ToyDown` (single-attribute, multi-attribute, and non-negative)  
298 on a simplified version of the same data 2010 data. We executed 16 runs of each variant  
299 with each of five different splits of the privacy budget across the five lower levels of the  
300 census geographic hierarchy (Table 1), fixing the total budget for those five levels at  $\varepsilon = 1$ .  
301 The data was derived from the reconstructed Texas data simplified to include only seven  
302 distinct types: one for the total Hispanic population and one for each of six subgroups of  
303 the non-Hispanic population based on race (White; Black; American Indian; Asian; Native  
304 Hawaiian/Pacific Islander; and Some Other Race or multiple races). Post-processing for single-  
305 attribute `ToyDown` was implemented in NumPy, while post-processing for multi-attribute  
306 and non-negative `ToyDown` used a Gurobi solver.

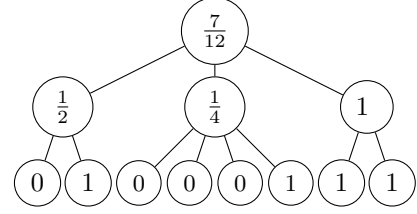
## 307 5 Hierarchical budget allocation

308 The relationship of the hierarchical allocation  $(\varepsilon_1, \dots, \varepsilon_d)$  to various measures of output  
309 accuracy is not obvious. On one hand, it might seem that higher values of  $\varepsilon_d$  (the block-level  
310 budget) will best promote accuracy at the block level, for a fixed  $\varepsilon$ . But on the other  
311 hand, imposing hierarchical consistency forces lower levels to be consistent with the totals at  
312 higher levels, which means that noise at higher levels can trickle down to lower levels. These  
313 competing effects create tradeoffs that are hard to balance without further analysis.

## 5:8 Census TopDown: The Impacts of Differential Privacy on Redistricting

Split name	state $\varepsilon_2$	county $\varepsilon_3$	tract $\varepsilon_4$	BG $\varepsilon_5$	block $\varepsilon_6$
equal	0.2	0.2	0.2	0.2	0.2
state-heavy	0.5	0.25	0.083	0.083	0.083
tract-heavy	0.083	0.167	0.5	0.167	0.083
BG-heavy	0.083	0.083	0.167	0.5	0.167
block-heavy	0.083	0.083	0.083	0.25	0.5

■ **Table 1** Names of designated budget splits used in **ToyDown** and **TopDown** runs below, each with a budget of  $\varepsilon_1 = 9$  on the nation and a total of 1 allocated below the national level.



■ **Figure 1** A district in a three-level hierarchy. The 0/1 weight of a leaf indicates its membership in the district; each non-leaf weight is the average of the node's children.

### 314 5.1 ToyDown error expressions

315 ► **Definition 1** (District, weights, error). A district  $D \subseteq H_d$  is a subset of the leaves (blocks)  
 316 of the hierarchy  $H$ . For hierarchy  $H$ , a district  $D$  induces weights  $w_h \in [0, 1]$  on the hierarchy  
 317 nodes, defined recursively as follows:

- 318 ■ For each leaf  $h \in H_d$ , let  $w_h = 1$  if  $h \in D$  and  $w_h = 0$  otherwise.
- 319 ■ For  $\ell \leq d - 1$  and  $h \in H_\ell$ , let  $w_h = \frac{1}{n(h)} \cdot \sum_{i \in [n(h)]} w_{h_i}$  be the average of the weights of  
 320 the children.

321 In a homogeneous hierarchy, we can observe that each  $w_h$  equals the fraction of the leaves  
 322 descended from  $h$  that belong to  $D$ . In particular, the root weight is  $w_1 = |D|/|H_d| = 1/k$  if  
 323 there are  $k$  districts of equal population made from nodes of equal population.

324 For node  $h \in H$ , we record the error  $E_h = \alpha_h - a_h$  introduced by **ToyDown** to the count  
 325  $a_h$ . The total error over district  $D$  is  $E_D = \sum_{h \in D} E_h$ . Let  $\hat{h}$  denote the parent of node  $h$ .

326 ► **Theorem 2** (Error expressions).  $E_1 = L_1$ . For  $\ell \in \{2, \dots, d\}$  and non-root node  $h_i \in H_\ell$ ,  
 327 and for every district  $D$  with associated weights  $w_h$  on the nodes,

$$328 \quad E_{h_i} = L_{h_i} + \frac{1}{n(h)} \left( E_h - \sum_{j \in [n(h)]} L_{h_j} \right), \quad E_D = w_1 L_1 + \sum_{h \in H \setminus \{1\}} (w_h - w_{\hat{h}}) L_h. \quad (1)$$

329 We make several observations. First, our intuition that error at higher levels trickles down  
 330 to lower levels is correct, but this effect is rather weak. The error at a child  $h_i$  is determined  
 331 by the parent error  $E_h$  discounted by the degree  $n(h)$ , the number of siblings. This suggests  
 332 that placing more budget at level  $\ell$  is an efficient way to secure accuracy at that level, until  
 333 a fairly extreme level of error at higher levels overwhelms the degree-based “discount.”

334 Second, because the  $L_h$  are all independent random variables with  $\mathbb{E}(L_h) = 0$  and  
 335  $\text{Var}(L_h) = 8/\varepsilon_{\ell(h)}^2$ , the theorem provides the following expression for variance that we use  
 336 repeatedly.

337 ► **Corollary 3** (Error expectation and variance). For all  $D \subseteq H_d$  and associated weights  $w_h$ ,  
 338 the expected error and error variance produced by **ToyDown** satisfy  $\mathbb{E}(E_D) = 0$  and

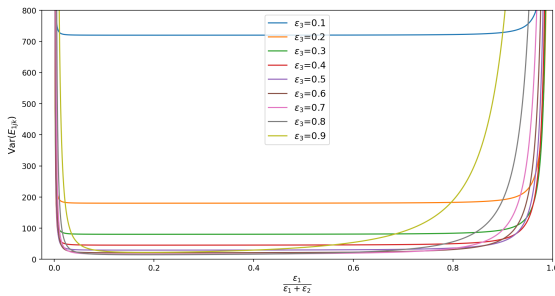
$$339 \quad \text{Var}(E_D) = \frac{8w_1^2}{\varepsilon_1^2} + \sum_{\ell=2}^d \left( \frac{8}{\varepsilon_\ell^2} \cdot \sum_{h \in H_\ell} (w_h - w_{\hat{h}})^2 \right). \quad (2)$$

340 Third, we get a more explicit expression if restricting to homogeneous hierarchies  $H$ .  
 341 Consider the case of a singleton district  $\{h\}$  made of a single census block  $h \in H_d$ .

342 ► **Corollary 4** (Error variance, homogeneous case). *The ToyDown error for a single block*  
 343  *$h \in H_d$  satisfies*

$$344 \quad \text{Var}(E_h) = \frac{8}{\varepsilon_1^2(n_1 \cdots n_{d-1})^2} + \sum_{\ell=2}^d \frac{8n_{\ell-1}(n_{\ell-1} - 1)}{\varepsilon_\ell^2(n_{\ell-1} \cdots n_{d-1})^2}. \quad (3)$$

345 Figure 2 plots this expression for various ways of splitting a total privacy budget of  
 346  $\varepsilon = 1$  across a three-level hierarchy with  $n_1 = n_2 = 10$ . The minimum of  $f(x_1, \dots, x_d) =$   
 347  $\sum_{\ell=1}^d a_\ell/x_\ell^2$  subject to  $\sum_\ell x_\ell = \varepsilon$  and  $x_\ell \geq 0$  is achieved at  $x_\ell = \varepsilon a_\ell^{1/3}/\sum_i a_i^{1/3}$  for all  $\ell$ . For  
 348 the example in Figure 2, the minimum-variance split is  $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.038, 0.171, 0.791)$  with  
 349 variance 14.52. (See accompanying [CoLab notebook](#).) One important note in interpreting  
 350 Figure 2 is that these variance numbers are absolute and don't depend on knowing population  
 351 counts for the nodes of the hierarchy. They are simply based on sampling Laplace noise with  
 352 the given parameters. If a variance of about 15 in the bottom-level counts is too high to be  
 353 tolerated in an application, one would have to increase  $\varepsilon$  to achieve lower variance.



$\varepsilon$	Allocation	$L^1$ error
1.0	(.16, .16, .16, .16, .16, .2)	0.03
1.0	(.2, .16, .16, .16, .16, .16)	0.03
1.0	(.1, .1, .1, .1, .1, .5)	0.02
1.0	(.02, .02, .02, .02, .02, .9)	0.03
1.0	(.66, .30, .01, .01, .01, .01)	0.09

■ **Figure 2** ToyDown error variance for a leaf node in the three-level hierarchy with  $n_1 = n_2 = 10$  and  $\varepsilon = 1$ . The curves show varying  $\varepsilon_3$  (colors) and the relative balance of  $\varepsilon_1$  and  $\varepsilon_2$  ( $x$ -axis).

■ **Table 2**  $L^1$  error measurements from selected TopDown runs on reconstructed Texas data. The allocation  $(\varepsilon_1, \dots, \varepsilon_6)$  goes from the nation  $\ell = 1$  down to census blocks at  $\ell = 6$ .

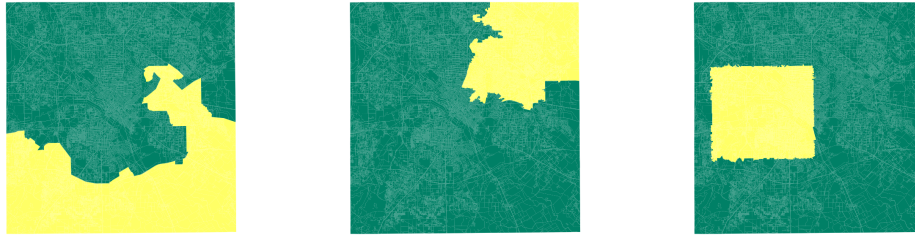
## 354 5.2 Empirical error experiments in TopDown

355 Next, we move to TopDown, which requires the use of input data. First, using reconstructed  
 356 2010 Texas data, we varied the relative allocation vector and the total  $\varepsilon$ , then measured  
 357 the effects with an  $L^1$  error metric included in the Census code [5]. This is a measure of  
 358 block-level error: it adds the magnitudes of changes in the bins, then divides by twice the  
 359 total population in the histogram.

360 Table 2 reports a small selection of the 100+ different scenarios explored. In general, the  
 361 lowest error outcomes were observed in a few scenarios: when the budget was distributed  
 362 near-equally to the levels of the hierarchy, and when half of the available budget was placed  
 363 at the bottom level—beyond  $\varepsilon_d = \varepsilon/2$ , further bottom-weighting gave diminishing returns in  
 364 block-level accuracy.

365 But a budget allocation that produces small block-level errors may not produce small  
 366 errors for *districts*, depending on the degree of cancellation or correlation. Next, we use  
 367 random district generation to understand the effects of off-spine aggregation. In particular,  
 368 we employ the Markov chain sampling algorithm called *recombination* (or ReCom), which runs  
 369 an elementary move that fuses two neighboring districts and re-partitions the double-district  
 370 by a random balanced cut to a random spanning tree [10].

## 5:10 Census TopDown: The Impacts of Differential Privacy on Redistricting



■ **Figure 3** Three sample districts (yellow) in Dallas County, each within two percent of the ideal population for  $k = 4$  districts. These are drawn by tract ReCom, block ReCom, and a square-favoring algorithm, respectively.

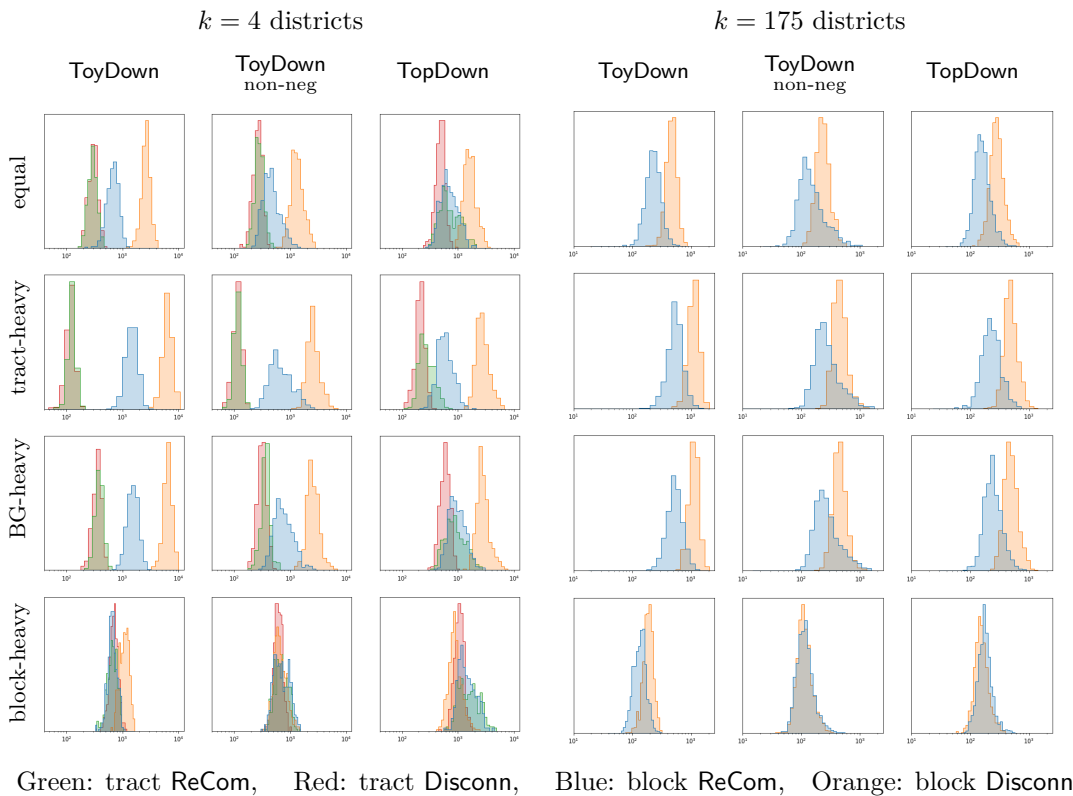
371 We begin with county commission districts in Dallas County, where  $k = 4$ . Since the 2010  
372 population of Dallas County was roughly 2.4 million, each district will have roughly 600,000  
373 people, making them nearly as big as congressional districts and much larger than tracts.  
374 We also include divisions of the county into  $k = 175$  districts of between 13,000 and 14,000  
375 people each for a small-district comparison. Figure 4 plots the data from our experiments on  
376 a logarithmic scale. Each histogram displays 400 values, one for each district drawn by the  
377 specified district-drawing algorithm; each value is the mean observed district-level population  
378 error magnitude over 16 executions of the specified hierarchical noising algorithm using the  
379 specified budget allocation.

380 First, consider two unrealistic forms of district-generation: tract Disconn (red) and block  
381 Disconn (orange), which randomly choose units of the specified type until assembling a  
382 collection with the appropriate population. These are unrealistic because they do not form  
383 connected districts; here, they are used to illustrate the effects of aggregation, neglecting  
384 spatial factors entirely. We see in Figure 4 that block-based methods generate hugely more  
385 error than tract-based methods, except if the budget allocation is concentrated at the bottom  
386 of the hierarchy. The effect is stronger for ToyDown (in keeping with Theorem 2), but is  
387 easily observed for TopDown as well.

388 We compare that with the more realistic district-generation algorithm block ReCom  
389 (blue), which builds compact and connected districts out of block units. This tends to give  
390 error levels in between the extremes set by the other two. Likewise, tract ReCom (green)  
391 builds compact and connected districts from tracts. One reasonable mechanism by which  
392 ReCom has much lower error than Disconn is that ReCom districts will tend to have higher  
393 “hierarchical integrity,” keeping higher-level units whole just by virtue of being connected  
394 and plump. The interior of ReCom districts will thus contain many whole block groups  
395 and tracts. Near the boundary, block groups and tracts are more fragmented, leaving the  
396 corresponding block-level errors uncanceled. These fragmentation ideas are explored more  
397 fully in Section 6 and some sample districts are depicted here.

398 The cancellation effect is significant: in most experiments, the error level for ReCom  
399 districts is much closer to that of tract Disconn than block Disconn (recall the data is plotted  
400 on a logarithmic scale). Overall, drawing districts out of larger pieces (e.g., using tract  
401 Disconn instead of ReCom, or ReCom instead of block Disconn) lowers error magnitude  
402 significantly in the best case and has little or no effect in the worst case.

403 Although tract ReCom and tract Disconn behave very similarly under ToyDown, the  
404 compact districts perform noticeably worse than their disconnected relatives once we pass  
405 to the full complexity of TopDown. At first this seems puzzling, because compact and



**Figure 4** These histograms show district-level error on a log scale for various combinations of budget splits (rows), district-drawing algorithms (colors), and noising algorithms (columns). We include both large districts and small districts, dividing the county into  $k = 4$  and  $k = 175$  equal parts. Each histogram displays 400 values, one for each district drawn by the specified algorithm, plotting the mean observed district-level population error magnitude over 16 executions of the noising algorithm using the specified budget allocation.

connected districts are being punished by the geography-aware TopDown. But the reason for this is apparent on further reflection: *spatial autocorrelation* is causing the post-processing corrections to move nearby tracts in the same direction, impeding the cancellation that makes counts usually more accurate on larger geographies.

In the end, the story that emerges from these investigations is that, with full TopDown, the best accuracy that can be observed for large districts occurs when they are made from whole tracts and the allocation is tract-heavy; an equal split is not much worse. For districts with population around 13,000,  $\epsilon = 1$  noising creates errors in the low hundreds for compact, connected districts, with the best performance for block-heavy allocations. Again, an equal split is not much worse, suggesting that this might be a good policy choice for accuracy in districts across many scales.

## 6 Geometrically compact vs hierarchically greedy districts

The analysis above suggests that the district-level error  $E_D$  will depend not only on the randomness of the noising algorithms, but also on the geometry of  $D$  and the structure of  $H$ . This section studies the hypothesis that districts that disrespect the geographical hierarchy will tend to have higher error magnitude. This section defines the *fragmentation score*,



## 5:12 Census TopDown: The Impacts of Differential Privacy on Redistricting

relates a district's fragmentation score to its error variance under `ToyDown`, and compares the fragmentation of two simple district-drawing algorithms on homogeneous hierarchies and simple geographies. Ultimately, we find that the explanatory value of the fragmentation score decays as we move to more realistic deployment of `TopDown`. This discrepancy raises important questions for future study: Which of the many additional features of `TopDown` attenuates the fragmentation–variance relationship?

We define a score intended to capture the contribution to  $\text{Var}(E_D)$  of the shape of the district with respect to the hierarchy. Recall that  $\hat{h}$  denotes the parent of node  $h$ .

► **Definition 5** (Fragmentation score). For  $D \subseteq H_d$ , let  $\text{Frag}(D) = \sum_{h \in H} (w_h - w_{\hat{h}})^2$ .

Because weights are in  $[0, 1]$ , the score obeys  $0 \leq \text{Frag}(D) < |H|$  for all districts, with higher scores indicating the presence of more units that are only partially included in  $D$ .

This fragmentation score is reverse-engineered from the expression for the variance of district-level population errors when using `ToyDown` with privacy divided equally across levels of the hierarchy (Corollary 3): namely,  $\text{Var}(E_D) = \frac{8d^2}{\varepsilon^2} (w_1^2 + \text{Frag}(D))$ . When the district  $D$  itself is a random variable sampled from some distribution, the expected fragmentation  $\mathbb{E}(\text{Frag}(D))$  is similarly related to  $\text{Var}(E_D)$ . Namely, using the law of total variation, when each level gets  $\varepsilon/d$  privacy budget:

$$\text{Var}(E_D) = \mathbb{E}(\text{Var}(E_D|D)) + \text{Var}(\mathbb{E}(E_D|D)) = \mathbb{E}(\text{Var}(E_D|D)) = \frac{8d^2}{\varepsilon^2} (\mathbb{E}(\text{Frag}(D)) + \mathbb{E}(w_1^2)).$$

When  $\varepsilon$  is allocated unequally across levels, as for the other splits in Table 1, the simple analytical relationship between the fragmentation score and the error variance breaks down.

Observe that a hierarchy  $H$  does not capture all of the geometry relevant to district drawing. In particular,  $H$  does not directly encode any information about block adjacency, and therefore we can't detect from  $H$  that a district is contiguous. For algorithms to generate contiguous districts, we need to make use of the plane geometry associated to  $H$ . We restrict our attention to the simplest case: homogeneous hierarchies (where every node on level  $\ell < d$  has exactly  $n_\ell$  children) and *square tilings*. (where each unit on level  $\ell$  is a square and has  $n_\ell$  children that cover it with a  $\sqrt{n_\ell} \times \sqrt{n_\ell}$  grid tiling).

We analyze the fragmentation score for two simple district-drawing algorithms (see Appendix C). The `Greedy` algorithm builds a district from the largest subtrees possible, only subdividing a subtree when necessary. It takes as input  $H$  and  $k \in \mathbb{N}$  and returns a district of size  $N = \lfloor |H_d|/k \rfloor$ , assembled by starting with the largest available units at random and adding units that are adjacent in the labeling sequence without passing size  $N$ , then allowing one partial unit, and so on recursively at lower levels. Observe that `Greedy` depends only on the hierarchy  $H$ . The `Square` algorithm takes as input a square, homogeneous hierarchy  $H$  and  $k \in \mathbb{N}$  such that the district size is a perfect square,  $|D| = |H_d|/k = s_d^2$ . It outputs a uniformly random  $s_d \times s_d$  square of blocks.

► **Theorem 6.** Let  $D_G \sim \text{Greedy}(H, k)$ ,  $D_\square \sim \text{Square}(H, k)$ . For  $n_1 \cdot n_2 \cdots n_{d-2} \geq k \geq 2$ , let  $L = \arg \min\{\ell : n_1 \cdot n_2 \cdots n_\ell \geq k\}$ .

$$\mathbb{E}(\text{Frag}(D_G)) \leq \frac{k-1}{k^2} \sum_{\ell=1}^L n_\ell + \frac{1}{4} \sum_{\ell=L+1}^{d-1} n_\ell; \quad \mathbb{E}(\text{Frag}(D_\square)) \geq \frac{2}{3} \left( \frac{\sqrt{n_1 \cdots n_{d-1}}}{\sqrt{k}} - \frac{11}{2} \right) \sqrt{n_{d-1}}.$$

Dallas County is nearly a perfect square shape, so it gives us an opportunity to set some roughly realistic parameters to evaluate these bounds. There are 529 tracts in Dallas County,

with an average of 3.2 blocks groups per tract and 26.4 blocks per block group, yielding 44,113 total blocks. We can approximate these parameters by setting  $d = 4$ , using  $k = 4$  as for the county commission districts, and setting  $(n_1, n_2, n_3) = (484, 4, 25)$  which has a reasonably similar 48,400 blocks (as a result,  $L = 1$ ). The bounds in the theorem say that  $\mathbb{E}(\text{Frag}(D_G)) \leq 98$  and  $\mathbb{E}(\text{Frag}(D_\square)) \geq 264$ . Note: for homogeneous hierarchies  $H$  with equal-population leaves, the score  $\text{Frag}(D_G)$  is independent of algorithm randomness and can be computed exactly; for the above parameters  $\text{Frag}(D_G) = 90.75$ . So the bound in the theorem is fairly tight, at least in this case.

To interpret the theorem, it is helpful to think of **Greedy** as being hierarchically greedy and **Square** as being geometrically greedy. That is, the former is oriented toward using the biggest possible units and keeping them whole, so that spatial considerations are secondary; the latter is oriented towards “compact” geographies with a lot of area relative to perimeter, and unit integrity is secondary. The theorem shows that compactness alone (a function of the plane geometry) does not keep down the fragmentation score (a function of the hierarchy), and indeed the bounds get farther apart as the hierarchy gets larger and more complicated. In Appendix C, we compare these theoretical results to empirical district errors, finding that fragmentation tracks well with errors in **ToyDown**, but that the complexity of the **TopDown** model weakens the relationship, suggesting a need for more sophisticated tools.

## 7 Ecological regression with noise

### 7.1 Inference methods for Voting Rights Act enforcement

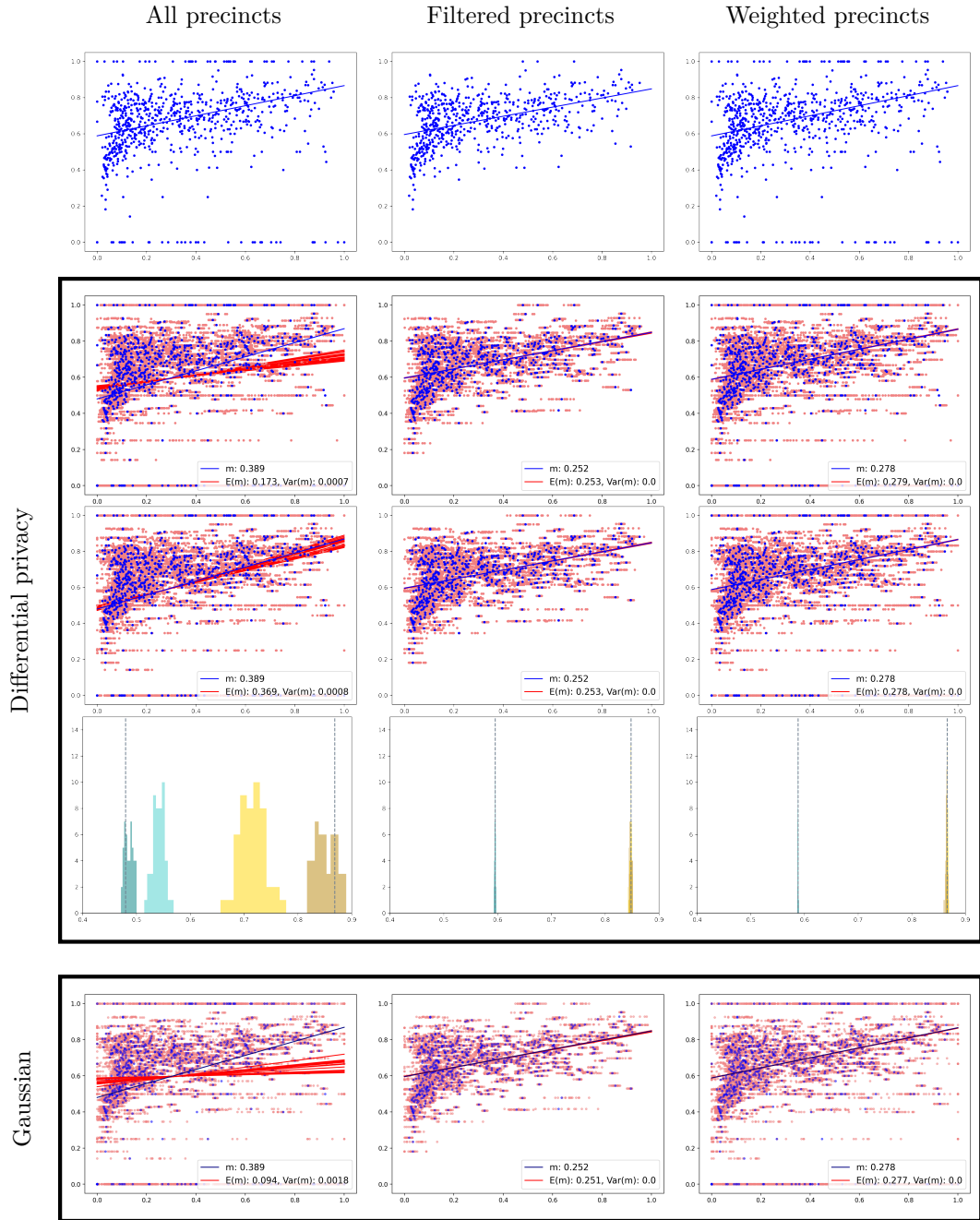
When elections are conducted by secret ballot, it is fundamentally impossible to precisely determine voting patterns by race from the reported outcomes alone. The standard methods for estimating these patterns use the cast votes at the precinct level, combined with the demographics by precinct, to infer racial polarization. Because the general aggregate-to-individual inference problem is called “ecological” (cf. ecological paradox, ecological fallacy), the leading techniques are called *ecological regression* (ER) and *ecological inference* (EI). It is rare that EI and ER do not substantively agree, and we focus on ER here because it lends itself to easily interpretable pictures.

ER is a simple linear regression, fitting a line to the data points determined by the precincts on a demographics-vs-votes plot. A high slope (positive or negative) indicates a likely strong difference in voting preferences, which is necessary to demonstrate the Gingles 2-3 tests for a VRA lawsuit.

The top row of Figure 5 shows standard ER run on the precincts of Dallas County, with each precinct plotted according to its percentage of Hispanic voting age population or HVAP ( $x$ -axis) and the share of cast votes that went to Lupe Valdez ( $y$ -axis). Strong racial polarization would show up as a fit line of high slope. This process produces a point estimate of Hispanic support for Valdez, found by intersecting the fit line with the  $x = 1$  line, which represents the scenario of 100% Hispanic population. The point estimate of non-Hispanic support for Valdez is at the intersection of the fit line with  $x = 0$ .

### 7.2 Summary of Experiments

**ToyDown** and **TopDown** were both run on the full Texas reconstruction from 2010. We plotted Dallas County votes from three contests: votes for Obama for president in 2012 general election, votes for Valdez for governor in the 2018 Democratic Party primary runoff, and votes for Chevalier for comptroller in the 2018 general election. We chose these contests



■ **Figure 5** Comparing ecological regression on un-noised data (top row) with various styles of noising. ER is re-run on data noised by differentially private ToyDown (second row), and data noised by TopDown (third row), both with  $\epsilon = 1$ , equal split. The blue dots repeat the un-noised data, the pink dots show 16 runs of noised data with pink fit lines re-computed each time. Below that, the histograms show the point estimates of Latino (gold) and non-Latino (teal) support for Valdez estimated from ER on data noised by ToyDown (lighter) and TopDown (darker). The last row contrasts the differentially private algorithms with a naive variant that adds noise to each precinct from a mean-zero Gaussian distribution, set to match the average precinct level  $L^1$  error observed in the ToyDown runs (in this case, this is  $\sigma = 20$ ). Across all of these experiments, the conclusion is striking: TopDown performs better than ToyDown and far better than a naive Gaussian variant, even without filtering precincts; if precincts are filtered or weighted, none of the noising alternatives threatens the ability to detect racially polarized voting.

Race	All precincts (827)		Filtered precincts (626)		Weighted precincts (827)	
	this group	complement	this group	complement	this group	complement
Hispanic	0.869	0.480	0.848	0.596	0.866	0.588
Black	0.917	0.518	0.851	0.620	0.835	0.595
White	0.555	0.623	0.474	0.811	0.478	0.805

Race	Algorithm	statistic	All (827)		Filtered (626)		Weighted (827)	
			group	compl.	group	compl.	group	compl.
Hispanic	ToyDown	mean	0.715	0.541	0.848	0.595	0.867	0.588
Hispanic	ToyDown	variance	36000	7000	250	43	160	19
Black	ToyDown	mean	0.798	0.543	0.851	0.62	0.835	0.595
Black	ToyDown	variance	39000	2100	89	5.9	25	2.1
White	ToyDown	mean	0.476	0.674	0.473	0.811	0.478	0.805
White	ToyDown	variance	17000	8000	64	36	33	17
Hispanic	TopDown	mean	0.853	0.485	0.848	0.595	0.865	0.587
Hispanic	TopDown	variance	45000	6700	480	100	120	16
Black	TopDown	mean	0.91	0.52	0.85	0.62	0.835	0.595
Black	TopDown	variance	30000	1200	250	23	45	2.4
White	TopDown	mean	0.582	0.607	0.472	0.81	0.47	0.804
White	TopDown	variance	10000	3400	92	37	92	10

■ **Table 3** Point estimates from ER for Dallas County in the Valdez/White primary runoff in 2018. In the first table, estimates are made with (un-noised) VAP data from the 2010 Census. In the *filtered precincts* case, precincts with fewer than 10 cast votes are excluded from the initial set of 827 precincts. In the *weighted precincts* case, precincts are weighted by the number of cast votes. The *ToyDown* and *TopDown* estimates are made from VAP data from 16 runs with  $\epsilon = 1$  and an  $\epsilon$ -budget with all levels given equal weighting. Variance is the empirical variance over the repeated runs of the noising algorithm and is in units of  $10^{-8}$ , shown to two significant digits.

500 because in each, ER finds evidence of strong racially polarized voting when using published  
 501 2010 census data. All three contests gave similar findings; we’ll choose the Valdez runoff  
 502 contest as our focus here.

503 For both *ToyDown* and *TopDown*, we vary how we handle the inclusion of small precincts in  
 504 the ecological regression. The options are All (every precinct is a data point in the scatterplot,  
 505 all weighted equally); Filtered (only including precincts with at least 10 votes cast in that  
 506 election); or Weighted (weighting the terms in the objective function in least-squares fit by  
 507 number of votes cast). Filtering and weighting are done using the exact number of cast votes,  
 508 not the differentially private precinct population totals, which is realistic to the use case.

509 For each noising run we have a block- or precinct-level matrix,  $\hat{M}$  of noised counts, with  
 510 height  $b$ , the number of geographic units (blocks or precincts), and width  $c$ , the number of  
 511 attributes for which there are counts recorded. We also have a corresponding matrix  $M$  of  
 512 un-noised counts. We can compute the  $L_1$  error by summing over the absolute value of every  
 513 entry in  $M - \hat{M}$ . *ToyDown* and *TopDown* were run 16 times for each configuration. Let  $E_{avg}$   
 514 be the average  $L_1$  error across noising runs.

515 If we add *Gaussian* noise to each count instead, the expected  $L_1$  error is  $\sum_{i,j} E[|X_{i,j}|]$ ,  
 516 where  $X_{i,j} \sim \mathcal{N}(0, \sigma^2)$ . This is the half-normal distribution, so  $E[|X_{i,j}|] = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$ . We  
 517 rearrange to find the standard deviation  $\sigma = \frac{E_{avg}\sqrt{\pi}}{bc\sqrt{2}}$  that defines the Gaussian distribution  
 518 (with  $\mu = 0$ ), so that adding a random variable drawn from it to each unit count will produce  
 519 an expected  $L^1$  error matching the average  $E_{avg}$  observed across the runs.

520 **7.3 The role of small precincts**

521 Practitioners who use ER have raised two questions regarding the effect of differential privacy:  
 522 (1) How robust will the estimate be after the noising? (2) Will noising diminish the estimate  
 523 of candidate support from a minority population? We analyzed the effects of **TopDown** and  
 524 **ToyDown** on the 2018 Texas Democratic primary runoff election, where Lupe Valdez was a  
 525 clear minority candidate of choice in Dallas county.<sup>1</sup>

526 We begin by observing that of the 827 precincts in Dallas County, 201 have fewer than  
 527 10 cast votes from that election day—in fact, 99 precincts recorded zero cast votes. These  
 528 precincts are a big driver of instability under DP. This is not surprising; percentage swings  
 529 are much higher in small numbers even if the noise injected might be low. However, down-  
 530 weighting these small precincts makes the estimate almost always agree with the un-noised  
 531 estimate. Specifically, we assign weights to the precincts equivalent to the number of total  
 532 votes in the precinct. Figure 5 shows how the estimates vary by run type and data treatment.

533 **8 Conclusion**

534 The central goal of this study has been to take the concerns of redistricting practitioners  
 535 seriously and to investigate potential destabilizing effects of **TopDown** on the status quo. A  
 536 second major goal is to make recommendations, both to the Disclosure Avoidance team at  
 537 the Census Bureau and to the same practitioners—the attorneys, experts, and redistricting  
 538 line-drawers in the field. Texas generally, and Dallas County in particular, was selected  
 539 because it has been the site of several interesting Voting Rights Act cases in the last 20  
 540 years.<sup>2</sup>

541 Our top-line conclusion is that, at least for the Texas localities and election data we  
 542 examined, **TopDown** performs far better than more naive noising in terms of preserving  
 543 accuracy and signal detection for election administration and voting rights law. Perhaps  
 544 more importantly, we have created an experimental apparatus to help other groups conduct  
 545 independent analyses.

546 This work has led us to isolate several elements of common redistricting practice that lead  
 547 to higher-variance outputs and more error under **TopDown**. The first example is the common  
 548 use of a full precinct dataset, with no population weighting, in running racial polarization  
 549 inference techniques. The second major example is the use of the smallest available units,  
 550 census blocks, for building districts of all sizes, with no particular priority on intactness  
 551 for larger units of Census geography. In both cases, we find that these were already likely  
 552 sources of silent error. Filtering small precincts (or, better, weighting by population) and  
 553 building districts that prioritize preserving whole the largest units that are suited to their  
 554 scale are two examples of simple updates to redistricting practice. Besides being sound on  
 555 first principles, these adjustments can insulate data users from DP-related distortions and  
 556 help safeguard the important work of fair redistricting.

---

<sup>1</sup> We also examined the general elections for President in 2012 and Comptroller in 2018, with similar findings.

<sup>2</sup> This is a large county with considerable racial and ethnic diversity. Follow-up work will consider smaller and more racially homogeneous localities.

## References

- 557 —
- 558 1 *13 U.S.C. Section 9*. URL: <https://www.law.cornell.edu/uscode/text/13/9>.
- 559 2 John Abowd, Daniel Kifer, Brett Moran, Robert Ashmead, Philip Leclerc, William  
560 Sexton, Simson Garfinkel, and Ashwin Machanavajjhala. Census topdown: Differentially  
561 private data, incremental schemas, and consistency with public knowledge. 2019. URL:  
562 [https://github.com/uscensusbureau/census2020-dase2e/blob/master/doc/20190711\\_](https://github.com/uscensusbureau/census2020-dase2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf)  
563 [0945\\_Consistency\\_for\\_Large\\_Scale\\_Differentially\\_Private\\_Histograms.pdf](https://github.com/uscensusbureau/census2020-dase2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf).
- 564 3 *Avery v. Midland County*, 390 U.S. 474 (1968).
- 565 4 U.S. Census Bureau. *Disclosure avoidance system - End to End demonstration*. URL:  
566 <https://github.com/uscensusbureau/census2020-das-e2e>.
- 567 5 U.S. Census Bureau. *Disclosure avoidance system - End to End demonstration,*  
568 *L1 metric*. URL: [https://github.com/uscensusbureau/census2020-das-e2e/blob/](https://github.com/uscensusbureau/census2020-das-e2e/blob/3f2c9cf9cb3c33a4e2067bd784ff381792f7ffc0/programs/validator.py#L20)  
569 [3f2c9cf9cb3c33a4e2067bd784ff381792f7ffc0/programs/validator.py#L20](https://github.com/uscensusbureau/census2020-das-e2e/blob/3f2c9cf9cb3c33a4e2067bd784ff381792f7ffc0/programs/validator.py#L20).
- 570 6 U.S. Census Bureau. *TopDown: Adding Geometric Noise to Counts*.  
571 URL: [https://github.com/uscensusbureau/census2020-das-e2e/blob/](https://github.com/uscensusbureau/census2020-das-e2e/blob/d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown_engine.py#L678)  
572 [d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown\\_engine.py#](https://github.com/uscensusbureau/census2020-das-e2e/blob/d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown_engine.py#L678)  
573 [L678](https://github.com/uscensusbureau/census2020-das-e2e/blob/d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown_engine.py#L678).
- 574 7 U.S. Census Bureau. *2010 Demonstration Data Products*, 2010. URL: [https://www.](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html)  
575 [census.gov/programs-surveys/decennial-census/2020-census/planning-management/](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html)  
576 [2020-census-data-products/2010-demonstration-data-products.html](https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html).
- 577 8 U.S. Census Bureau. *2010 Census Summary File 1*, 2012. URL: [https://www.census.gov/](https://www.census.gov/prod/cen2010/doc/sf1.pdf)  
578 [prod/cen2010/doc/sf1.pdf](https://www.census.gov/prod/cen2010/doc/sf1.pdf).
- 579 9 U.S. Census Bureau. *Census P.L. 94-171 Redistricting Data*, 2017. URL: [https://www.census.](https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html)  
580 [gov/programs-surveys/decennial-census/about/rdo/summary-files.html](https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html).
- 581 10 Daryl DeFord, Moon Duchin, and Justin Solomon. Recombination: A family of markov chains  
582 for redistricting. *arXiv preprint arXiv:1911.05725*, 2019.
- 583 11 Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings*  
584 *of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database*  
585 *systems*, pages 202–210, 2003.
- 586 12 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to  
587 sensitivity in private data analysis. *Halevi S., Rabin T. (eds) Theory of Cryptography. TCC*  
588 *2006. Lecture Notes in Computer Science*, 3876, 2006.
- 589 13 Peter Wayner JN Matthews, Bhushan Suwal. *Accompanying GitHub repository*. URL: [https:](https://github.com/nggg/census-diff-privacy)  
590 [//github.com/nggg/census-diff-privacy](https://github.com/nggg/census-diff-privacy).
- 591 14 Denis Kazakov. *Census Scripts GitHub repository*, 2019. URL: [https://github.com/](https://github.com/94kazakov/census_scripts)  
592 [94kazakov/census\\_scripts](https://github.com/94kazakov/census_scripts).
- 593 15 U.S. Census Bureau Michael Hawes. *Differential Privacy and the 2020 Decennial Census*, 2020.  
594 URL: [https://www2.census.gov/about/policies/2020-03-05-differential-privacy.](https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf)  
595 [pdf](https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf).
- 596 16 National Conference of State Legislatures. *2010 Redistricting Deviation Table*. URL: [https:](https://www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.aspx)  
597 [//www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.](https://www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.aspx)  
598 [aspx](https://www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.aspx).
- 599 17 *Reynolds v. Sims*, 377 U.S. 533 (1964).
- 600 18 *Wesberry v. Sanders*, 376 U.S. 1 (1964).

## 601 **A** ToyDown and TopDown

602 ToyDown is described in Algorithm 2. It uses the *Laplace distribution*  $\text{Lap}(b)$  with scale  
 603 parameter  $b$ , i.e., the probability distribution over  $\mathbb{R}$  with mean zero and probability density  
 604 function  $\mathbb{P}[L] = \frac{1}{2b}e^{-|L|/b}$ . It has variance  $2b^2$ . TopDown uses the *geometric* distribution, a  
 605 discretized version of the Laplace distribution with integer support.

606 The inputs to TopDown are as follows.  $A_{H,T} = \{a_{h,t}\}_{h \in H, t \in T}$ , where  $a_{h,t}$  is the number  
 607 of people in  $h$  of type  $t$ ;  $W = (Q_1, \dots, Q_{|W|})$  is a *workload* consisting of a collection of  
 608 histograms  $Q$ ;  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)$  is a hierarchical allocation of the privacy budget, with  $\varepsilon_\ell > 0$   
 609 at each level;  $B : W \rightarrow [0, 1]$  with  $\sum_{Q \in W} B(Q) = 1$  is a probability vector describing the  
 610 relative privacy budget on each histogram in the workload; *invariants*  $V$ ; and *structural*  
 611 *inequalities*  $S$ . We write  $\mathbf{a}_h = \{a_{h,t}\}_{t \in T}$  (and  $\boldsymbol{\alpha}_h$  analogously). For a query  $q$ , we write  
 612  $q(\mathbf{a}_h) = \sum_{t \in q} a_{h,t}$  (and  $q(\boldsymbol{\alpha}_h)$  analogously).

613 In the first stage (lines 2-5), a geometric random variable is added to the raw counts  $a$  to  
 614 produce noised counts  $\hat{a}$ . In the second stage (lines 6-8), the noised counts are adapted to  
 615 the nearest integer values that meet a collection of equality and inequality conditions. These  
 616 equalities and inequalities, over the real numbers, describe a convex polytope; therefore the  
 617 post-processing can be thought of geometrically as a closest-point projection to the integer  
 618 points in the convex body under  $L^2$  distance.

619 The noising stages of both ToyDown and TopDown are  $\varepsilon$ -differentially private for  $\varepsilon =$   
 620  $\sum_{\ell=1}^d \varepsilon_\ell$ . In ToyDown, this stage can be viewed as generating a single histogram at each  
 621 level  $\ell$  using budget  $\varepsilon_\ell$ . Following the Census Bureau, we use bounded differential privacy,  
 622 wherein the global sensitivity of histogram queries is 2. In TopDown, the budget at level  
 623  $\ell$  is further divided among the  $|W|$  histograms  $Q$  in the workload, each receiving  $B(Q)\varepsilon_\ell$   
 624 of the budget. Because ToyDown's post-processing is data independent, ToyDown is  $\varepsilon$ -DP.  
 625 TopDown's post-processing is not data independent: the invariants and structural inequalities  
 626 may depend on the original data.

### Algorithm 1 TopDown, based on [2]

---

```

1: procedure TOPDOWN( $A_{H,T}, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_d, W, B, V, S$ )
2:   for  $h \in H, Q \in W, q \in Q$  do
3:      $\beta \leftarrow \exp(-B(Q) \cdot \varepsilon_{\ell(h)}/2)$ 
4:      $G_{h,q} \leftarrow \text{Geom}(\beta)$  ▷ See [6]
5:      $\hat{a}_{h,q} \leftarrow q(\mathbf{a}_h) + G_{h,q}$  ▷ Geometric mechanism with
sensitivity 2, budget  $B(Q) \cdot \varepsilon_{\ell(h)}$ 

6:   for  $\ell = 1, \dots, d$  do
7:     Compute hierarchically-consistent ▷ A sophisticated heuristic algorithm
       non-negative integers  $\{\alpha_{h,t}\}_{h \in H_\ell, t \in T}$  out of scope for this work
       minimizing  $\sum_{h \in H_\ell} \sum_{q \in W_\ell} (q(\boldsymbol{\alpha}_h) - \hat{a}_{h,q})^2$ ,
       subject to the invariants:  $v^*(\boldsymbol{\alpha}_h) = v^*(\mathbf{a}_h)$  for all  $h \in H_\ell, v \in V$ 
       and structural inequalities:  $s(\boldsymbol{\alpha}_h, \mathbf{a}_h) \leq 0$  for all  $h \in H_\ell, s \in S$ 

8:   return  $\{\alpha_{h,t}\}_{h \in H, t \in T}$ 

```

---

---

**Algorithm 2** ToyDown

---

```

1: procedure TOYDOWN( $A_H = \{a_h\}_{h \in H}, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_d$ ) ▷ (Single attribute)
2:   for  $h \in H$  do
3:      $L_h \sim \text{Lap}(2/\varepsilon_{\ell(h)})$ 
4:      $\hat{a}_h \leftarrow a_h + L_h$  ▷ Laplace mechanism with sensitivity 2, budget  $\varepsilon_{\ell(h)}$ 
5:   for  $\ell = 1, \dots, d$  do
6:     Compute hierarchically consistent  $\{\alpha_h\}_{h \in H_\ell}$ 
       minimizing  $\sum_{h \in H_\ell} (\alpha_h - \hat{a}_h)^2$ 
7:   return  $\{\alpha_h\}_{h \in H}$ 

8: procedure MultiAttrTOYDOWN( $A_{H,T} = \{a_{h,t}\}_{h \in H, t \in T}, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_d$ )
9:   for  $h \in H, t \in T$  do
10:     $L_{h,t} \sim \text{Lap}(2/\varepsilon_{\ell(h)})$ 
11:     $\hat{a}_{h,t} \leftarrow a_{h,t} + L_{h,t}$  ▷ Laplace mechanism with sensitivity 2, budget  $\varepsilon_{\ell(h)}$ 
12:   for  $\ell = 1, \dots, d$  do
13:     Compute hierarchically consistent
       (optionally, non-negative)  $\{\alpha_{h,t}\}_{h \in H_\ell, t \in T}$ 
       minimizing  $\sum_{h \in H_\ell, t \in T} (\alpha_{h,t} - \hat{a}_{h,t})^2$ 
14:   return  $\{\alpha_{h,t}\}_{h \in H, t \in T}$ 

```

---

## B Detailed materials and methods

### B.1 Primary data sources

2010 US Census demographic data was downloaded using the Census API, and the 2010 census block, block group, and tract shapefile for Dallas County were downloaded from the US Census Bureau’s TIGER/Line Shapefiles. For our VRA analysis, we obtained both statewide election results and a statewide precinct shapefile from the Texas Capitol Data Portal, which we then trimmed to the precincts within Dallas County.<sup>3</sup>

We use a person-level dataset obtained by applying a reconstruction technique to the block-level data from Texas from the 2010 Census.<sup>4</sup> The reconstructed microdata records contain block-level sex, age, ethnicity, and race information consistent with a collection of tables from 2010 Census Summary File 1. We note that this reconstruction follows the same strategy used by the Census Bureau itself as the first step of its reidentification experiment [15], based on [11].

The reconstructed data is far from perfect. Unlike the Bureau, we do not have access to the ground truth data needed to quantify the errors. The Bureau’s own reconstruction experiment reconstructed 46% of entries exactly, plus an additional 25% within  $\pm 1$  year error in age [15]. We note that our reconstructed data contains no household information, because this was not present in the tables used in the constraint system. This is significant because the TopDown configurations for the US Census Bureau’s 2010 Demonstration Data Products [7] include household-based workload queries and invariants.

---

<sup>3</sup> Data comes from [data.capitol.texas.gov/topic/elections](http://data.capitol.texas.gov/topic/elections) and [data.capitol.texas.gov/topic/geography](http://data.capitol.texas.gov/topic/geography).

<sup>4</sup> A team led by data scientist and journalist Mark Hansen at Columbia, including Denis Kazakov, Timothy Donald Jones, and William Reed Palmer, designed an algorithm to solve for the detailed data, which we describe in this section. Code is available upon request [14].



647 **B.2 TopDown configuration**

648 The exact configuration files and code for all the runs are available in this paper’s accompanying  
 649 repository [13]. The TopDown code used for this paper was modified from the publicly  
 650 available demonstration release of the US Census Bureau’s Disclosure Avoidance System  
 651 2018 End-to-End test release [4]. The input data fed to the algorithm was obtained by  
 652 restructuring the reconstructed 2010 block-level Texas microdata into the 1940s IPUMs  
 653 data format. Most importantly, the reconstructions allowed for 63 distinct combination of  
 654 races whereas the End-to-End release only allows for 6 races, so all multi-racial entries were  
 655 re-categorized as Other in our TopDown runs.

656 Because TopDown’s post-processing is done level by level, the noisy counts in Dallas  
 657 County do not depend on the noisy counts at the tract-level or below in counties other than  
 658 Dallas. We modified the census reconstructed data to focus on Dallas county and minimize  
 659 the computation time spent processing the other 253 counties in Texas. Specifically, for every  
 660 non-Dallas county, we placed all of the population into a single block.

661 We do not enforce certain household invariants that the Census Bureau is planning to  
 662 enforce, and our workload omits household queries that are used in Census’s demonstration  
 663 data products. Our choice to omit household queries and invariants is result of our use of  
 664 reconstructed 2010 census microdata which does not include household information. We  
 665 did perform additional runs with household invariants and queries using crude synthetic  
 666 household data, the results of which are available in the data repository accompanying this  
 667 paper [13]. In those runs, the population in each block was grouped into households of size 5  
 668 with at most one group smaller than 5. Ultimately, we focused on the experiments that did  
 669 not require synthetic household data.

670 The TopDown runs without the household workload or invariants use a workload consisting  
 671 of two histograms:  $Q_{detailed}$  and  $Q_{va,eth,race}$  with 10% and 90% of the budget respectively.  
 672 (The additional runs with households includes an additional households and group quarters  
 673 histogram in the workload assigned 22.5% of the budget, leaving 10% and 67.5% for  $Q_{detailed}$   
 674 and  $Q_{va,eth,race}$  respectively.) The End-to-End TopDown code reports a differentially private  
 675 estimate of the  $L^1$  error with  $\varepsilon = 0.0001$  not included in privacy budget specified elsewhere  
 676 in the configuration file and discussed elsewhere in this paper.

677 **C District fragmentation****Algorithm 3** Greedy

---

```

1: procedure GREEDY( $H, k$ )
2:   if  $k = 1$  then
3:     Return  $H$ 
4:    $N \leftarrow \lfloor |H_d|/k \rfloor$ ,  $D \leftarrow \emptyset$ ,  $h^* \leftarrow h_1$ 
5:   while  $N > 0$  do
6:     For  $h^*$  and  $D$ , let  $S(h^*, D)$  be the set of
       children  $h$  of  $h^*$  that are disjoint from  $D$ .
7:     while  $\exists h \in S(h^*, D) : |h| \leq N$  do
8:        $D \leftarrow D \cup h$             $\triangleright$  Associating  $h$  with the blocks descendent from it
9:        $N \leftarrow N - |h|$ 
10:    Pick  $h^* \in S(h^*, D)$ 
return  $D$ 

```

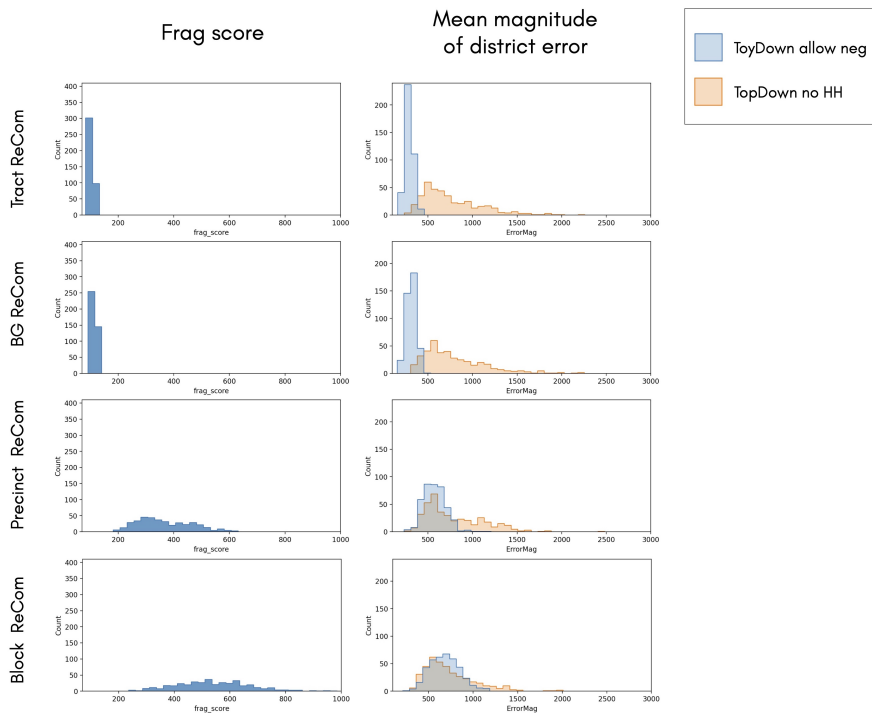
---

■ **Algorithm 4** Square

- 
- 1: **procedure** SQUARE( $H, k$ )
  - 2:    $s_d \leftarrow \sqrt{|H_d|/k}$  ▷ Side length in blocks of the district
  - 3:    $S_d \leftarrow \sqrt{n_1 \cdot n_2 \cdots n_{d-1}}$  ▷ Side length in blocks of the region
  - 4:   Sample  $i, j \in \{1, \dots, S_d - s_d + 1\}$  uniformly at random
  - 5:   **return**  $D_{i,j}$ , the square district with top left corner at  $(i, j)$
- 

678 In Section 6, we defined the fragmentation score and its relationship to error variance for  
 679 ToyDown, and analyzed the expected fragmentation score of districts produced by different  
 680 district drawing algorithms. Now we apply TopDown to examine the relationship between a  
 681 district’s population error and geometry, as captured by the fragmentation score.

682 We fix the a total budget and an equal allocation across levels:  $0.2 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = \varepsilon_5 =$   
 683  $\varepsilon_6$ , as in Table 1. (We do not need to noise the nation because we are focusing on Texas; we  
 684 do need to noise Texas even though its total population is invariant, because its population  
 685 by race is allowed to vary.) We apply ReCom to build districts out of tracts, block groups,  
 686 and blocks—all of which are part of the census hierarchy—and add a realistic variant that  
 687 builds from whole *precincts*. These are about the same size as block groups and are more  
 688 commonly used in redistricting.



■ **Figure 6** Do the building-block units of districts matter? Histograms of fragmentation score (left column) and mean error magnitude (right column) are shown across four district-drawing algorithms that prioritize compactness. (Dallas County,  $k = 4$ .) We see that using larger units leads to significantly lower fragmentation and correspondingly low district-level error in ToyDown, but the advantage erodes when we pass to TopDown.

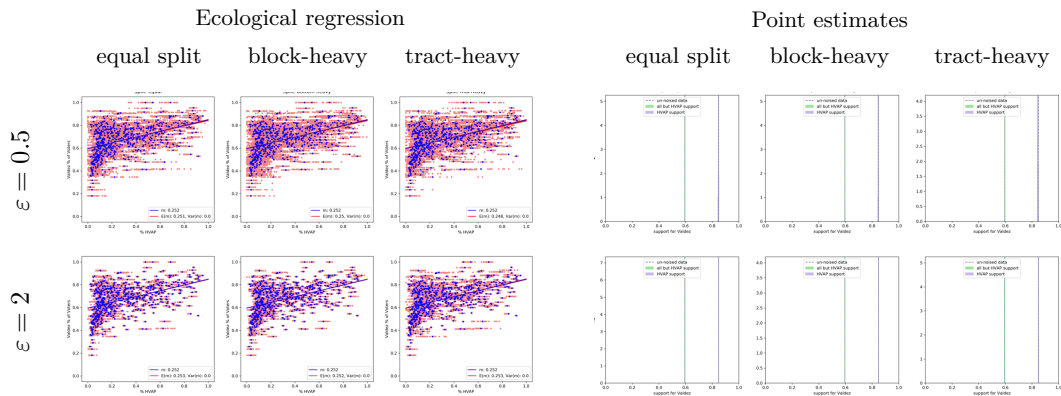
5:22 **Census TopDown: The Impacts of Differential Privacy on Redistricting**

689 Figure 6 plots the data from our experiments. Each of the 12 histograms displays 400  
 690 values, one for each district drawn by the specified district-drawing algorithm. The histograms  
 691 on the left plot the fragmentation score of each district; the histograms on the right plot the  
 692 mean observed district-level population error magnitude over 16 executions of the specified  
 693 hierarchical noising algorithm.

694 The size of the constituent units is observed to have a controlling effect on the fragmentation  
 695 score, as expected. As we would expect, this carries over to the simplest ToyDown (allowing  
 696 negativity). (Note that since the error has zero mean, higher variance drives up the mean  
 697 magnitude of error.) But the choice of base units makes far less difference by the time we  
 698 move to full TopDown. These observations are consistent, again, with a strong similarity  
 699 across spatially nearby units. All four kinds of ReCom will tend to produce compact, squat  
 700 districts whose units are more closely geographically proximal than would be observed with  
 701 disconnected or elongated shapes. Random noise is uncorrelated, but the post-processing  
 702 effects can be highly spatially correlated because of spatial relationships in the underlying  
 703 counts by race, ethnicity, and voting age.

704 **D Robustness of noisy ER**

705 Figure 7 extends the findings from Figure 5 with more splits and allocations, showing that  
 706 as long as small precincts are filtered out, ecological regression for RPV analysis in Dallas  
 707 County is robust to changes in the allocation of the privacy budget across the levels of the  
 708 hierarchy and the total privacy budget for TopDown. The corresponding plots for ToyDown  
 709 are essentially indistinguishable. (ER with precincts weighted by population is similarly  
 710 robust.)



**Figure 7** Ecological regression for the Valdez-White runoff election with  $\epsilon = .5$  and  $\epsilon = 2$  and three different budget allocations, together with corresponding point estimates for Latino and non-Latino support for Valdez, with small precincts filtered out as in Figure 5. Findings stay remarkably stable.

---

# Formal Privacy Methods for the 2020 Census

---

Contact: Gordon Long — [glong@mitre.org](mailto:glong@mitre.org)

April 2020

JSR-19-2F

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

JASON  
The MITRE Corporation  
7515 Colshire Drive  
McLean, Virginia 22102-7508  
(703) 983-6997



**IRC\_00949**



---

# Contents

<b>1</b>	<b>EXECUTIVE SUMMARY</b>	<b>1</b>
1.1	Findings . . . . .	6
1.1.1	The re-identification vulnerability . . . . .	6
1.1.2	The use of Differential Privacy . . . . .	6
1.1.3	Stakeholder response . . . . .	7
1.1.4	The pace of introduction of Differential Privacy . . . . .	7
1.2	Recommendations . . . . .	7
1.2.1	The re-identification vulnerability . . . . .	7
1.2.2	Communication with external stakeholders . . . . .	8
1.2.3	Deployment of Differential Privacy for the 2020 census and beyond . . . . .	8
<b>2</b>	<b>INTRODUCTION</b>	<b>11</b>
2.1	Overview of the Census . . . . .	11
2.2	Overview of the Study . . . . .	13
2.3	Overview of the Report . . . . .	13
<b>3</b>	<b>CENSUS PROCESS</b>	<b>17</b>
3.1	Census Geographical Hierarchy . . . . .	17
3.2	Census Process and Products . . . . .	21
3.3	The Need for Disclosure Avoidance . . . . .	26
<b>4</b>	<b>THE CENSUS RE-IDENTIFICATION VULNERABILITY</b>	<b>29</b>
4.1	Reconstruction of Census Tabular Data . . . . .	29
4.2	Results of Dinur and Nissim . . . . .	33
4.3	JASON Verification of the Dinur-Nissim Results . . . . .	34
4.4	Queries in the Presence of Noise . . . . .	38
4.5	Information Theory and Database Uniqueness . . . . .	40
<b>5</b>	<b>DIFFERENTIAL PRIVACY</b>	<b>43</b>
5.1	Mechanisms . . . . .	47
5.1.1	Laplace mechanism . . . . .	47
5.1.2	Geometric mechanism . . . . .	48
5.1.3	Matrix mechanism . . . . .	49
5.2	Some Surprising Results in Applying Differential Privacy . . . . .	50
5.2.1	Cumulative distribution functions . . . . .	50

---

5.2.2	Median . . . . .	51
5.2.3	Common mechanisms can give strange results for small n	53
5.2.4	Nearly equivalent queries with vastly different results . . .	55
5.3	Invariants . . . . .	55
5.4	Database Joins under Differential Privacy . . . . .	57
5.5	The Dinur-Nissim Database under Differential Privacy . . . . .	58
5.6	Multiple Query Vulnerability . . . . .	60
5.7	Disclosure Avoidance using Differential Privacy . . . . .	62
<b>6</b>	<b>ASSESSING THE ACCURACY-PRIVACY TRADE-OFF</b>	<b>69</b>
6.1	Census Analysis of 2010 Census Data . . . . .	69
6.2	IPUMS Analysis of 1940 Census Data under the Census DAS . . .	72
<b>7</b>	<b>MANAGING THE TRADE-OFF OF ACCURACY, GRANULARITY AND PRIVACY</b>	<b>81</b>
7.1	Risk Assessment . . . . .	82
7.2	Engaging the User Community . . . . .	83
7.3	Possible Impacts on Redistricting . . . . .	85
7.4	Limiting Release of Small Scale Data . . . . .	86
7.5	The Need for Special Channels . . . . .	86
<b>8</b>	<b>Conclusion</b>	<b>89</b>
8.1	The Census Vulnerability Raises Real Privacy Issues . . . . .	89
8.2	Two Statutory Requirements are in Tension in Title 13 . . . . .	92
8.3	Findings . . . . .	94
8.3.1	The re-identification vulnerability . . . . .	94
8.3.2	The use of Differential Privacy . . . . .	95
8.3.3	Stakeholder response . . . . .	96
8.3.4	The pace of introduction of Differential Privacy . . . . .	96
8.4	Recommendations . . . . .	97
8.4.1	The re-identification vulnerability . . . . .	97
8.4.2	Communication with external stakeholders . . . . .	97
8.4.3	Deployment of Differential Privacy for the 2020 census and beyond . . . . .	98
<b>A</b>	<b>APPENDIX: Information Theory and Database Uniqueness</b>	<b>99</b>
A.1	Noiseless Reconstruction via Linear Algebra . . . . .	99
A.2	Information: An Introductory Example . . . . .	101

---

A.3	Information Gained Per Query . . . . .	103
A.4	Information Gained from Multiple Noiseless Queries . . . . .	104
A.5	$m$ Sequences and Hadamard Matrices . . . . .	107
A.6	The Minimal Number of Queries . . . . .	108
A.7	Noisy Single Queries . . . . .	109
A.8	Multiple Noisy Queries . . . . .	114
A.9	Reconstruction . . . . .	115
<b>B</b>	<b>MATLAB CODE FOR DN DATABASE RECONSTRUCTION</b>	<b>119</b>





---

## Abstract

In preparation for the 2020 decennial census, the Census Bureau asked JASON to examine the scientific validity of the vulnerability that the Census Bureau discovered in its traditional approach to Disclosure Avoidance, the methods used to protect the confidentiality of respondent data. To address this vulnerability, the Census Bureau will employ differential privacy, a mathematically rigorous formal approach to managing disclosure risk. JASON judges that the analysis of the vulnerability performed by Census is scientifically valid. The use of Differential Privacy in protecting respondent data leads to the need to balance statistical accuracy with privacy loss. JASON discusses this trade-off and provides suggestions for its management.



---

# 1 EXECUTIVE SUMMARY

A decennial population census of the United States will officially begin April 1, 2020. Under Title 13 of the US Code, the Bureau of the Census is legally obligated to protect the confidentiality of all establishments and individuals who participate in providing census data. In particular, Census cannot publish any information that could be used to identify a participant.

Over the years, a large amount of personal data have become easily available via online and commercial resources. It has also become much easier to analyze large amounts of data using modern computers and data-science tools. This has made it possible to breach the confidentiality protection promised to respondents of studies and surveys. There have been several notable examples in which records collected under pledges of confidentiality from a survey were linked with public data resulting in the re-identification of the individuals participating in the survey. In an exercise to evaluate the confidentiality protection of the census, the Census Bureau discovered such a vulnerability exists for their data as well.

Using the individual responses from participants (known as microdata), the Census Bureau produces a collection of tables that summarize population counts, age distributions, etc., for various levels of geographic resolution from the whole nation down to census blocks. A variety of approaches have been used by Census in the past to prevent re-identification. In addition to the removal of direct identifiers, Census applies geographic thresholding, top and bottom coding, swapping and other methods of obfuscation to hide identifying characteristics. It was previously thought to be computationally intractable to reconstruct the microdata from the published tabular summaries. But in 2018, applying modern optimization methods along with relatively modest computational resources, Census succeeded in reconstructing, from the published 2010 census data, geographic location (census block), sex, age, and ethnicity for 46% of the US population (142 million people). By linking the reconstructed microdata with information in commercial

---

databases, Census was then able to match and putatively re-identify 45% of the reconstructed records. Of these putative re-identifications, 38% were confirmed. This corresponds to 17% of the US population in 2010 (a total of over 52 million people). Such a re-identification rate exceeds that obtained in a previous internal Census assessment by four orders of magnitude. Public release of these re-identifications would constitute a substantial abrogation of the Census' Title 13 confidentiality obligations.

In view of these developments, Census has proposed the application of formal privacy methods, in particular, the use of Differential Privacy (DP). DP, introduced in 2006, has as its goal the prevention of learning about the participation of an individual in a survey by adding tailored noise to the result of any query on data associated with that survey. DP provides a set of algorithms used to compute statistical information about a dataset (e.g. counts, histograms, etc.), but infuses those statistics with tailored noise, making it possible to publish information about a survey while limiting the possibility of disclosure of detailed private information about survey participants.

A number of features make DP an attractive approach for protection of confidentiality for the 2020 census and beyond. Notably, privacy loss (in a technical sense) can be rigorously quantified via a privacy-loss parameter. In addition, there are techniques to create synthetic data such that subsequent queries will not cause further confidentiality loss provided such queries do not access the original data. Finally, confidentiality would degrade in a controlled way should it prove necessary to re-access the original data in order to publish further tabulations. Census proposes to use this approach by adding noise to the tabular summaries it traditionally publishes and then using these to reconstruct synthetic census microdata. Both the noised tabular summaries and the synthetic microdata could then be publicly released.

Once the differentially private tabulations and the synthetic data are produced, the use of DP methods offers a mathematically rigorous guarantee that any

---

further analysis of the released data preserves the original level of confidentiality protection. However, one drawback of such approaches is that the applied noise will degrade the accuracy of various tabulations and statistical analyses of the data, particularly those associated with small populations. Census data are used by a large number of government, academic, business, and other stakeholders. Census is therefore compelled to make an explicit trade-off between the accuracy of its data releases and the privacy of respondents.

Census charged JASON with the following three tasks:

1. Examine the scientific validity of the vulnerability that the Census Bureau discovered in the methods that it has historically used to protect the confidentiality of respondent data when preparing publications;
2. Evaluate whether the Census Bureau has properly assessed the vulnerability as described above;
3. Provide suggestions to represent the trade-offs between privacy-loss and accuracy to explicitly represent user choices.

JASON has not attempted to duplicate the reconstruction of census micro-data as it does not have access to that data, nor to data from commercial marketing databases. JASON has, however, confirmed via database simulation that such an attack is possible; JASON concludes that, provided one publishes a sufficient number of tabular summaries, there are multiple approaches using modern optimization algorithms to reconstruct the database from which the summaries originated with high probability. This creates a significant risk of disclosure of census data protected under Title 13.

Census plans to release some data without noise, most importantly, state populations for the apportionment of Congressional representatives. In addition, Public Law 94-171 requires that Census provide the states with small-area data necessary to perform legislative redistricting for both Federal and State electoral

---

districts. The Census has set up a voluntary program in which state officials define the geographic areas for which they wish to receive census data. While only population data are legally mandated, Census has traditionally also provided other demographic data such as race, ethnicity and voting age populations. For expedience, states have simply asked for these data at the finest geographical resolution (census blocks) and have then used the block populations to infer population counts for larger geographical areas such as legislative districts. The proposed creation of differentially private census tabulations will result in block-level populations that differ from the original census enumeration due to the infused noise. Releases of exact counts (known as invariants) are technically violations of DP in principle and degrade the privacy guarantee, although to what extent in practice remains a research issue. There arises, then, a tension between the obligations under PL 94-171 to release population data for legislative purposes and the requirements of Title 13 to protect confidentiality.

For large populations, for example at the national, state, or even in many cases the county level, using DP does not unduly perturb the accuracy of statistical queries on the data provided the privacy-loss parameter is not set too low (implying the infusion of a large amount of noise). This is important for diverse users of census data (demographers, city planners, businesses, social scientists etc.). But as the size of the population under consideration becomes smaller, the contributions from injected noise will more strongly affect such queries. Note that this is precisely what one wants for confidentiality protection, but is not desirable for computation of statistics for small populations. Thus there is also a tension between the need to protect confidentiality and the aim to provide quality statistical data to stakeholders. While the latter is not legally mandated for Census, it is aligned with the Office of Management and Budget's policy directive to agencies that produce useful governmental statistics, and Census has traditionally been a key supplier of such data through its various published products.

The trade-off between confidentiality and statistical accuracy is reflected in the choice of the DP privacy-loss parameter. A low value increases the level of

---

injected noise (and thus also confidentiality) but degrades statistical calculations. Another factor that also influences the choice of privacy-loss parameter is the number and geographical resolution of the tables released. For example, if no block-level data were publicly released, a re-identification “attack” of the sort described above presumably would become more difficult, perhaps making it feasible to add less noise and thus publish tables at a higher value of the privacy loss parameter than what would be required if block level tables were published. A re-identification attack, of the sort that originally led to the conclusions that more rigorous and effective confidentiality protections were required, has not been performed on microdata reconstructed from differentially private tabulations. Such an analysis is needed to gauge the level of protection needed.

Depending on the ultimate level of privacy protection that is applied for the 2020 census, some stakeholders may well need access to more accurate data. A benefit of differential privacy is that products can be released at various levels of protection depending on the level of statistical accuracy. The privacy-loss parameter can be viewed as a type of adjustable knob by which higher settings lead to less protection but more accuracy. However, products publicly released with too low a level of protection will again raise the risk of re-identification. One approach is to use technology (e.g. virtual machines, secure computation platforms etc.) to provide protected data enclaves that allow access to census data at lower levels of privacy protection to trusted stakeholders. Inappropriate disclosure of such data could still be legally enjoined via the use of binding non-disclosure agreements such as those currently in Title 13. This idea is similar to the concept of “need to know” used in environments handling classified information.

Finally, it will be necessary to engage and educate the various communities of stakeholders so that they can fully understand the implications (and the need for) DP. These engagements should be two-way conversations so that the Census Bureau can understand the breadth of requirements for census data, and stakeholders can in turn more fully appreciate the need for confidentiality protection in the present era of “big data”, and perhaps also be reassured that their statistical



---

needs can still be met.

## **1.1 Findings**

### **1.1.1 The re-identification vulnerability**

- The Census has demonstrated the re-identification of individuals using the published 2010 census tables.
- Approaches to disclosure avoidance such as swapping and top and bottom coding applied at the level used in the 2010 census are insufficient to prevent re-identification given the ability to perform database reconstruction and the availability of external data.

### **1.1.2 The use of Differential Privacy**

- The proposed use by Census of Differential Privacy to prevent re-identification is promising, but there is as yet no clear picture of how much noise is required to adequately protect census respondents. The appropriate risk assessments have not been performed.
- The Census has not fully identified or prioritized the queries that will be optimized for accuracy under Differential Privacy.
- At some proposed levels of confidentiality protection, and especially for small populations, census block-level data become noisy and lose statistical utility.
- Currently, Differential Privacy implementations do not provide uncertainty estimates for census queries.

---

### **1.1.3 Stakeholder response**

- Census has not adequately engaged their stakeholder communities regarding the implications of Differential Privacy for confidentiality protection and statistical utility.
- Release of block-level data aggravates the tension between confidentiality protection and data utility.
- Regarding statistical utility, because the use of Differential Privacy is new and state-of-the-art, it is not yet clear to the community of external stakeholders what the overall impact will be.

### **1.1.4 The pace of introduction of Differential Privacy**

- The use of Differential Privacy may bring into conflict two statutory responsibilities of Census, namely reporting of voting district populations and prevention of re-identification.
- The public, and many specialized constituencies, expect from government a measured pace of change, allowing them to adjust to change without excessive dislocation.

## **1.2 Recommendations**

### **1.2.1 The re-identification vulnerability**

- Use substantially equivalent methodologies as employed on the 2010 census data coupled with probabilistic record linkage to assess re-identification risk as a function of the privacy-loss parameter.
- Evaluate the trade-offs between re-identification risk and data utility arising from publishing fewer tables (e.g. none at the block-level) but at larger values of the privacy-loss parameter.

---

### **1.2.2 Communication with external stakeholders**

- Develop and circulate a list of frequently asked questions for the various stakeholder communities.
- Organize a set of workshops wherein users of census data can work with differentially private 2010 census data at various levels of confidentiality protection. Ensure all user communities are represented.
- Develop a set of 2010 tabulations and microdata at differing values of the privacy-loss parameter and make those available to stakeholders so that they can perform relevant queries to assess utility and also provide input into the query optimization process.
- Develop effective communication for groups of stakeholders regarding the impact of Differential Privacy on their uses for census data.
- Develop and provide to users error estimates for queries on data filtered through Differential Privacy.

### **1.2.3 Deployment of Differential Privacy for the 2020 census and beyond**

- In addition to the use of Differential Privacy, at whatever level of confidentiality protection is ultimately chosen, apply swapping as performed for the 2010 census so that no unexpected weakness of Differential Privacy as applied can result in a 2020 census with less protection than 2010.
- Defer the choice of the privacy-loss parameter and allocation of the detailed privacy budget for the 2020 census until the re-identification risk is assessed and the impact on external users is understood.
- Develop an approach, using real or virtual data enclaves, to facilitate access by trusted users of census data with a larger privacy-loss budget than those released publicly.

- 
- Forgo any public release of block-level data and reallocate that part of the privacy-loss budget to higher geographic levels.
  - Amid increasing demands for more granular data and in the face of conflicting statutory requirements, seek clarity on legal obligations for protection of data.



---

## 2 INTRODUCTION

### 2.1 Overview of the Census

The US decennial census, codified in law through the US Constitution has taken place every 10 years since 1790. The 24th such census will take place in 2020. The authority to collect and analyze the information gathered by the Census Bureau originates in Title 13 of the US Code enacted into law in 1954. Title 13 Section 9 of the US code mandates that neither the Secretary of Commerce or any other employee or officer of the Dept. of Commerce may

“... use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Dept or bureau or agency thereof to examine the individual reports.”

Census employees are sworn to uphold the tenets of Title 13 and there are strict penalties including fines and imprisonment should there be any violation. To ensure the mandate of Title 13 is upheld, the Census has traditionally used what are termed Disclosure Avoidance techniques on its publicly released statistical products. The particular approaches used by the Census for Disclosure Avoidance have evolved over the years. A short overview is contained in this report.

Surveys have long been an invaluable tool in determining policy and in the performance of social science and demographic research. In many cases such surveys require respondents to reveal sensitive information under the promise that such information will remain confidential. Traditionally, protection from disclosure was accomplished by anonymizing records. In this way, statistical analyses on issues of public importance could be accomplished while protecting the identity of the respondent. Over time however, the availability of public external data

---

and the increase in capability of data analytics has made protecting confidential data a challenge. By linking information in one data set with that of another containing some intersecting information (known as a record-linkage attack) it is sometimes possible to connect an anonymous record containing confidential information with a public record and thus identify the respondent. This is called re-identification of previously de-identified data. A number of newsworthy re-identifications have been accomplished in this way. Several approaches have been put forth to make such record linkage attacks harder (see e.g., [32]) but to date none of these have proven to be sufficiently robust to attack.

In 2016, analysts at the Census realized that, even though the Census publishes for the most part tabular summaries of its surveys, enough information could be gleaned from the results to correctly reconstruct a substantial fraction of the detailed survey responses. By linking this information with commercial marketing databases, the names of the respondents could be ascertained, a putative violation of Title 13.

In response, Census has proposed to utilize methods of formal privacy developed and analyzed in the cryptography community; Census proposes to use the methods of Differential Privacy (DP) [8] to secure the 2020 Census. Census requested a JASON study as part of the process of verifying their assessment of disclosure risk as well as assessing the proposed use of formal privacy approaches. Census' charge to JASON was as follows:

- JASON will examine the scientific validity of the vulnerability that the Census Bureau discovered in the methods it has historically used to protect the confidentiality of respondent data when preparing publications.
- Risk assessment: has the Census Bureau properly assessed the vulnerability?
- Implementing formal privacy requires making explicit choices between the accuracy of publications and their associated privacy loss; users always

---

want more accuracy, but the Census Bureau must also safeguard the respondents' privacy. How do we represent the trade-offs between privacy loss and accuracy to explicitly represent user choices? Are there other conceptual approaches we should try?

## **2.2 Overview of the Study**

JASON was introduced to the relevant issues through a set of presentations listed in Table 2-1. The briefers were experts both internal and external to the Census Bureau in areas such as disclosure avoidance, demography, and applications of census data such as redistricting. These talks were of high quality and were instrumental in educating JASON on these issues. In addition, members of JASON participating in the study were sworn into Title 13 allowing them to be briefed on information protected under this statute and providing JASON with important insights into the details of 2020 Census and particularly the Disclosure Avoidance system based on DP proposed for 2020. Finally, Census provided with JASON with a rich set of reference materials, some protected under Title 13. Details associated with those materials protected under Title 13 are not included in this report.

## **2.3 Overview of the Report**

In Section 3, we provide a brief overview of the census process, the information that Census is mandated to provide and the associated timeline. We also briefly review the methods that were used for Disclosure Avoidance in the past. In Section 4, we review the work that led Census to conclude that the previous approaches to Disclosure Avoidance were inadequate given the increasing availability of large datasets of personal information. In this context, we discuss the seminal work of Dinur and Nissim [5] leading to what is now called the Fundamental Law of Information Recovery. We also describe some experiments asso-



Speaker	Title	Affiliation
Ron Jarmin	Overview of the Dual Mandate and Legal and Historical Background for Disclosure Avoidance	US Census
Victoria Velkoff	Proposed 2020 Census Data Products	US Census
James Whitehorne	Overview of Redistricting Data Products	US Census
John Abowd	The Vulnerability in the 2010 Census Disclosure Avoidance System (DAS)	US Census
Ashwin Machanavajjhala	Interpreting Differential Privacy	Duke University
Dan Kifer	Design Principles of the TopDown Algorithm	Penn State University
Phil Leclerc	Empirical Analysis of Utility-Privacy Trade-offs for the TopDown Algorithm	US Census
William Sexton	Disclosure Avoidance At-Scale and Other Outstanding Issues	US Census
Cynthia Hollingsworth	How 2020 Census Data Products are Prepared	US Census
Rachel Marks	How 2020 Census Data Products Reflect Data User Feedback	US Census
Ken Hodges	How 2020 Census Products will be used by Demographers	Claritas
Justin Levitt	Uses of 2020 Census Redistricting Data	Loyola University
Tommy Wright	Suitability Assessment of Data Treated by DA Methods for Redistricting	US Census
Kamaliika Chaudhuri	Formal Privacy and User-Imposed Constraints	UCSD
Salil Vadhan	Formal Privacy and Data Analysis, Including Invariants	Harvard
Dave van Riper	Differential Privacy and the Decennial Census (via VTC)	U. Minnesota
Danah Boyd	Video Teleconference	Microsoft
Jerry Reiter	Video Teleconference	Duke University

Table 2-1: Briefers for JASON Census study.

ciated with the Dinur-Nissim work that underscore the conclusions of that work. In Section 5, we describe briefly the proposed use of DP as a means of protecting sensitive Census data. DP grew out of the work described above by Dinur and Nissim and then extended by Dwork and her collaborators [7]. DP makes possible statistical queries regarding a dataset to be performed while offering a rigorous bound on the amount one learns about a dataset if one record is deleted, added or replaced. Note that this is not, strictly speaking a guarantee of disclosure avoidance but it does provide in a rigorous way the likelihood of a record linkage attack. It does this by adding specially calibrated noise to the result of a specific query made on the dataset. For queries that involve large populations, the addition of noise does not unduly perturb the statistical accuracy of the query. But as a query focuses on smaller and smaller populations the noise will make it increas-

---

ingly difficult to infer individual characteristics. An attractive feature of DP is that the level of protection is tunable via the setting of a privacy loss parameter. The value set for the privacy loss parameter is meant to be a policy decision.

In Section 6, we discuss the results of some of the early work performed by Census on applying DP to census data. Census proposes to use DP to process the sensitive microdata and create the standard tabular summaries. Noise will then be added to these summaries to make them differentially private. The assessment of the privacy loss budget to be used has not yet been performed. Census will then use the same reconstruction algorithms it applied on the 2010 census data on the noised tables. This will create synthetic microdata that, in principle, should be safe to publish openly. We discuss some early applications of this approach and the nature of the synthetic data it produces. The proposed use of DP will lead to tension between protecting privacy while providing accurate demographic data for activities like redistricting. In Section 7 we propose some approaches for managing this trade-off. Finally in Section 8 we summarize our findings and recommendations.



---

## 3 CENSUS PROCESS

In this section we provide a brief overview of the main products that the Census provides as well as the geographic hierarchy that Census has established to collect the relevant respondent data. We also cover the approach the Census has used to process and summarize the required data. Finally, we discuss the evolving need for preservation of the confidentiality of Census data.

### 3.1 Census Geographical Hierarchy

The Census organizes the US population via a geographic hierarchy shown in Figure 3-1. At the top of this hierarchy are the national boundaries of the United States and Puerto Rico. Within each state, Census further subdivides the population according to county of residence. Counties are then further divided into tracts, block groups, and finally the lowest gradation of Census geography, the Census block. Census also surveys the households in each block and counts for example the number of residents, whether the resident owns or rents etc. Census also collects data for what are known as Group Quarters. Examples of these are dormitories, prisons, etc. The designations in Figure 3-1 of nation, region, state, county, tracts, block groups, and finally census blocks is called the "central spine" of the census geographic hierarchy. Off this spine are also indicated other important state and local divisions. For these, Census provides geographies that can then be used to determine counts in these regions off the spine. These Census geographies inform the placement of Census blocks so that the counts in these areas can be performed from Census block data.

The distribution of population and the number of households in a census tract, block group or block varies greatly across the nation. A map of the population density from 2010 census data is shown in Figure 3-2. As can be seen, the population density varies from thousands of people per square mile as for example in areas like New York City or Los Angeles down to less than ten people per

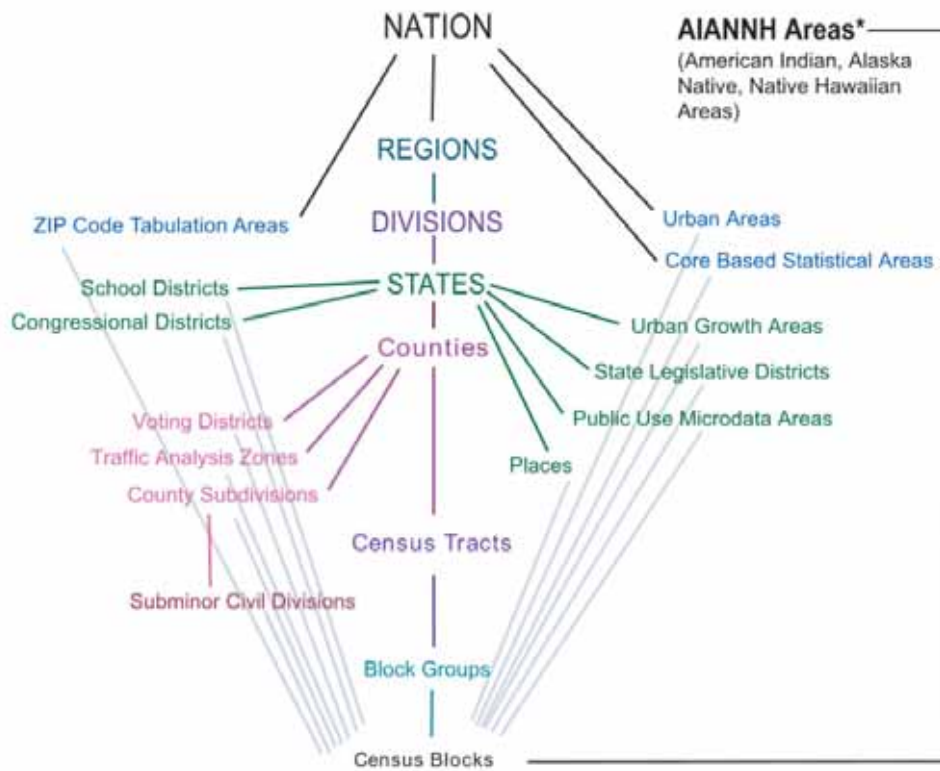


Figure 3-1: The geographical hierarchy used by Census in organizing its various surveys [38].

square mile in states such as Nevada. This diversity in the number of residents and number of households in various regions is one of the reasons Census must work to protect respondent information. In many cases, because of the uniqueness of a given area, it may be possible to identify census respondents. For example, in Figure 3-3 we display graphical representations of the distribution of population and number of households for the country in the form of Violin plots. As can be seen, there is wide variability in both population and number of households even at the census block level. Census blocks are comprised for the most part of roughly several hundred people, but in densely populated areas there are outliers with several thousand people; there is a similar picture for the number of households in a block. Block groups are larger consisting of typically a few thousand

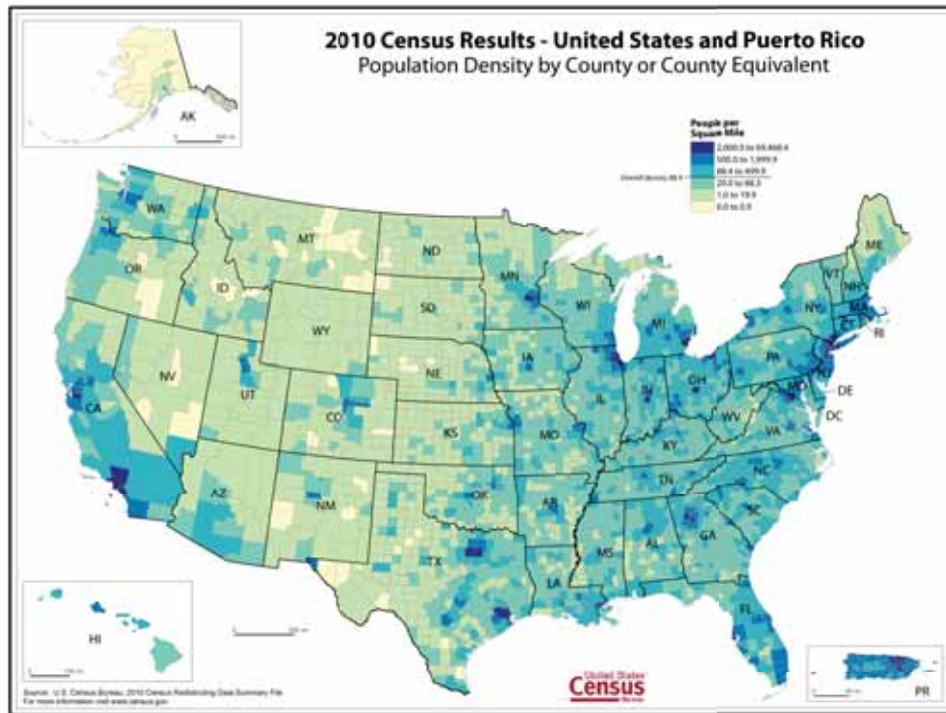


Figure 3-2: Map of population density across the United States from the 2010 census [35].

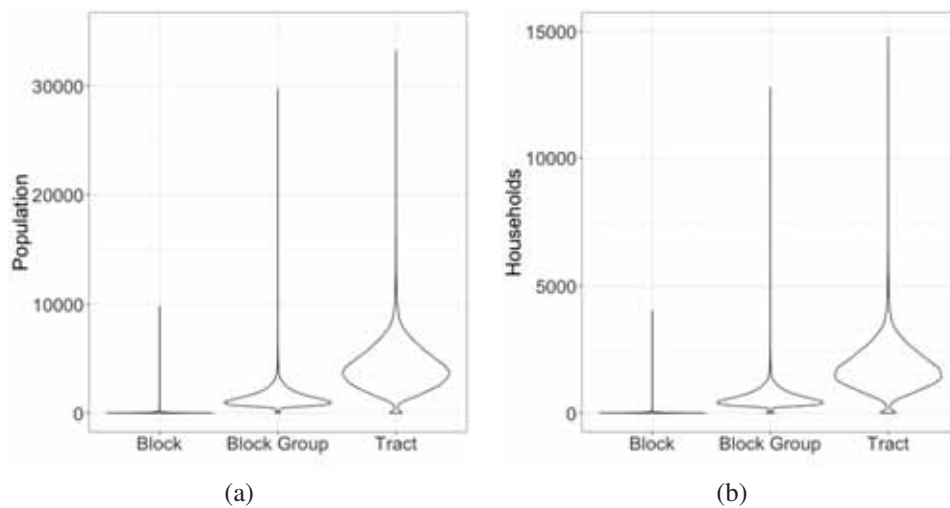


Figure 3-3: Violin plots of population and households for census tracts, block groups and blocks across the nation.

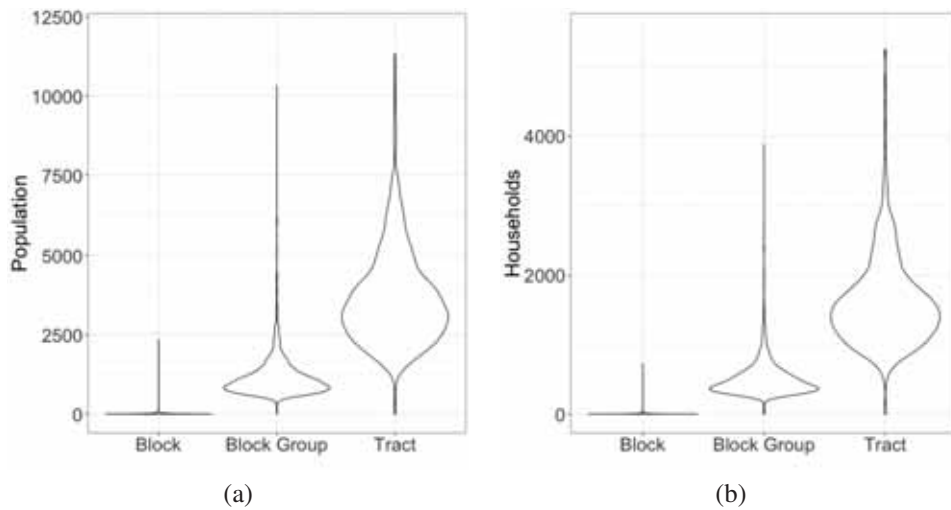


Figure 3-4: Violin plots of population and households for census tracts, block groups and blocks in Iowa.

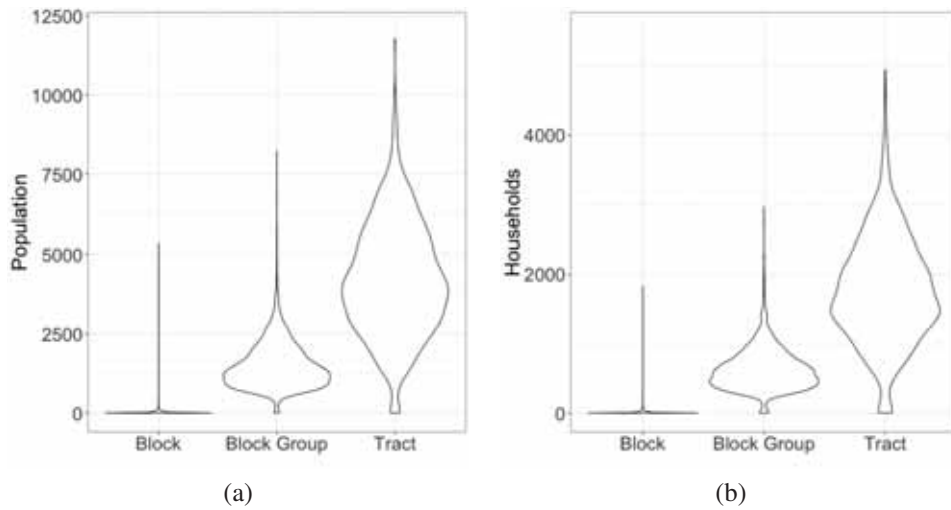


Figure 3-5: Violin plots of population and households for census tracts, block groups and blocks in Virginia.

people, but here also there is considerable variability. Census tracts may range from population sizes of several hundred in very sparsely populated areas to upwards of 30,000 people. The distribution of population and number of households for blocks, block groups and tracts in a state like Iowa is shown in Figure 3-4. This should be contrasted with the distribution for Virginia shown in Figure 3-5.

---

Finally it is important to note that census blocks do not always line up with other regions of interest. An important example is the use of census data to determine boundaries of both Congressional and State Legislative districts. Shown in Figure 3-6 are the boundaries for two Congressional districts in Virginia. The boundaries for the districts are shown in black. Census tracts are indicated in purple; census block groups are indicated in orange; and census blocks are indicated in gray. The boundaries for tracts, groups and blocks are quite complex indicative of geography but also complex population patterns. The boundaries of a Congressional district (as well as a state legislative district) are determined through a redistricting process that makes use of the information provided in the PL94 census product (discussed below).

### **3.2 Census Process and Products**

By April 1, 2020 (Census Day) every home will receive a request to participate in the 2020 census. This is the reference data for which respondents report where they usually live. Census then also canvasses group quarters (dorms, etc.) in April. Respondents indicate

- The number of people who live and sleep in a residence most of the time; the homeless are asked to respond as well,
- The ownership status of the household,
- Sex of the residents of the household,
- Age of the residents and their date of birth,
- Whether the residents are of Hispanic origin,<sup>1</sup>
- Race of the residents. This can be any or all of the 63 possible races as designated by the Office of Management and Budget (OMB).

---

<sup>1</sup>Census refers to this information as the Hispanicity of the respondent.



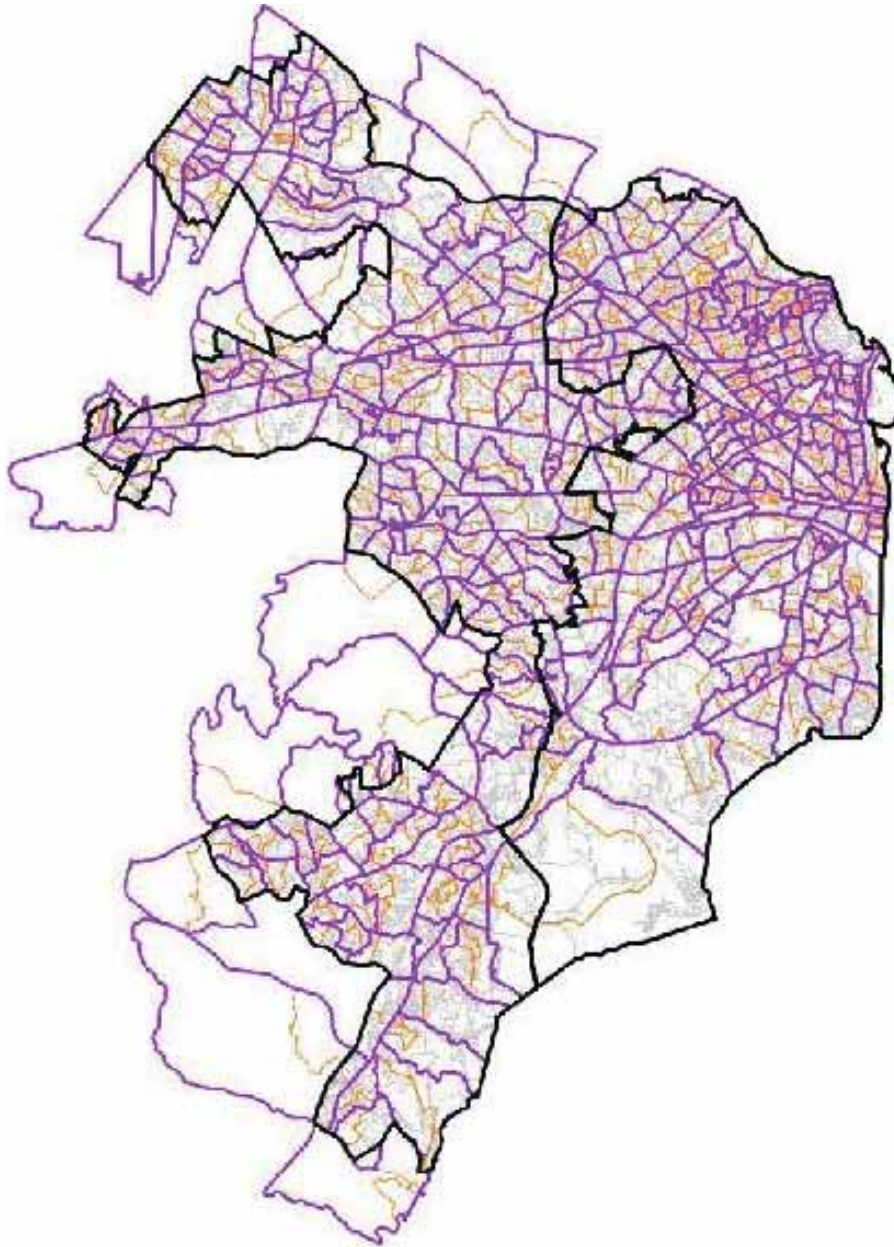


Figure 3-6: A map of two adjoining Congressional districts in Virginia. The black lines indicate the district boundaries; the purple lines indicate boundaries of census tracts; the orange lines indicate boundaries of block groups; the gray lines indicate census blocks.

---

The 2020 census will also collect information about US citizenship, but respondents will not be asked to indicate their citizenship on the census questionnaire. Instead this will be inferred from existing administrative records (e.g. Social Security Administration, Internal Revenue Service, etc.).

The respondent data are collected into a set of what Census terms microdata, a list of records indicating the responses for each resident. As the responses are received, records are de-duplicated and addresses are validated to insure that every person is counted only once. This forms the Census Unedited File or CUF. Where data are missing or inconsistent the Census employs a process known as imputation and edits the CUF to produce the hundred percent detail file or HDF. The final step is to identify those cells in the various tabular summaries where it may be possible to identify respondents. Here the Census performs confidentiality edits and swaps households as discussed further in Section 3.3. From here the various tabular summaries would be produced.

The Census Bureau through its surveys is responsible for the following products:

**Apportionment count** Apportionment is the process of dividing the 435 seats of the House of Representative among the states. The count is based on the resident population (both citizen and non-citizen) of the 50 states. An example of the result from the 2010 Census is shown in Figure 3-7 and must be delivered to the President and Congress by December, 2020.

**PL94-171** Public law 94-171 directs the Census Bureau to provide redistricting data for the 50 states. This is the first product that must be produced after the apportionment count is complete. Within a year of the 2020 census, the Bureau must send data agreed-upon with the states to redraw state congressional and legislative districts. To meet this requirement the Census has set up a voluntary program that makes it possible for states to receive population estimates as well as racial and Hispanicity distributions for areas relevant to the state congressional and legislative election process. An example of the tables provided in this product is shown

STATE	APPORTIONMENT POPULATION (APRIL 1, 2010)	NUMBER OF APPORTIONED REPRESENTATIVES BASED ON 2010 CENSUS	CHANGE IN SEATS FROM CENSUS 2000 APPORTIONMENT
Alabama	4,802,982	7	0
Alaska	721,523	1	0
Arizona	6,412,700	9	+1
Arkansas	2,926,229	4	0
California	37,341,989	53	0
Colorado	5,044,930	7	0
Connecticut	3,581,628	5	0
Delaware	900,877	1	0
Florida	18,900,773	27	+2
Georgia	9,727,566	14	+1
Hawaii	1,366,862	2	0
Idaho	1,573,499	2	0
Illinois	12,864,380	18	-1
Indiana	6,501,582	9	0
Iowa	3,053,787	4	-1
Kansas	2,863,813	4	0
Kentucky	4,350,606	6	0
Louisiana	4,553,962	6	-1
Maine	1,333,074	2	0

Source: 2010 Census Apportionment, Table 1

Figure 3-7: A partial list of the apportionment count determining the number of Congressional representatives from each state [39].

in Figure 3-8.

**Summary File 1** Census produces a set of demographic profiles after the apportionment and redistricting reports are complete. Summary File 1 (SF1) provides population counts for the 63 OMB race categories and Hispanicity down to the census block level. The report contains data from questions asked of all people and about every housing unit and includes sex, age, race etc. The report consists of 177 population tables, 58 housing tables down to the block level as well as tabulations at the county and tract level. SF1 also provides special tabulations for areas such as metropolitan regions, Congressional districts, school districts etc.

**Summary File 2** Summary File 2 (SF2) contains cross-tabulations of information on age, sex, household type, relationship, size for various races as well as Hispanicity down to census tract level as long as the population in the tract exceeds 100 people.

	Virginia	Block 1000, Block Group 1, Census Tract 2001.02, Alexandria city, Virginia	Block 1001, Block Group 1, Census Tract 2001.02, Alexandria city, Virginia
Total:	8,001,024	0	658
Population of one race:	7,767,624	0	630
White alone	5,486,852	0	225
Black or African American alone	1,551,399	0	176
American Indian and Alaska Native alone	29,225	0	3
Asian alone	439,890	0	132
Native Hawaiian and Other Pacific Islander alone	5,980	0	0
Some Other Race alone	254,278	0	94
Two or More Races:	233,400	0	28
Population of two races:	214,276	0	27
White; Black or African American	62,204	0	1
White; American Indian and Alaska Native	25,771	0	0
White; Asian	59,051	0	3
White; Native Hawaiian and Other Pacific Islander	2,618	0	1

Figure 3-8: An example of a population table in the PL94-171 summary file [39].

**American Community Survey** The American Community Survey (ACS) is an ongoing survey that has taken the place of the decennial long form. It is performed annually. Each year Census contacts 3.5 million households and asks that they fill out a detailed questionnaire. The survey is far more extensive than the decennial census and gathers information about household makeup, type of housing, citi-

---

zenship, employment etc. The information is used by a variety of stakeholders. Perhaps most importantly, the data are used to guide the disbursement of federal and state funds.

**Public Use Microdata Sample** Census provides edited samples of the micro-data records that make up the decennial census and the ACS. These records are assembled for areas that contain a minimum population of 100,000 (known as PUMAs) and are edited to protect confidentiality. The PUMS provides only a 10% sample of a PUMA.

### 3.3 The Need for Disclosure Avoidance

It was realized early on that some disclosure avoidance was necessary as the population and housing densities of the United States are not distributed in a homogeneous manner. Owing to special aspects of a location it may be possible to identify the particular person or persons living there. This would constitute a violation of Title 13. For example, Liberty Island, the base of the Statue of Liberty has one household listed, that of the Superintendent of the Monument and his wife [13]. Thus by focusing on this location and using external sources it should be possible to identify the residents of that particular household. For this reason, the information for this location is swapped with that of another household. A history of the methods used in the past 50 years to effect disclosure avoidance is available in the paper by McKenna [24]. We briefly describe these here to provide some context for this report. The discussion below is not complete but illustrates the evolution of the need to offer improved disclosure avoidance.

**Long form data** Long form census data have never been published at the lowest level of census geography (presently census blocks). The long form data were generally collected as part of the decennial census but in 2010 this data was relegated to what is now called the American Community Survey (ACS) which began

---

in 2005. The ACS only publishes data down to the block group level.

**1970 Census** The 1970 Census utilized suppression of whole tables as opposed to suppression of cells. The choice to suppress was based on the number of people in households in a given area. This approach had limitations in that tables with complementary information were not suppressed making it possible in some cases to infer the suppressed information. As indicated by McKenna, cells within an original table could still show an estimate of 1 or 2 people.

**1980 Census** The 1980 Census retained the approach of the 1970 census but modified it further by now suppressing tables with complementary information and zeroing cells with counts of 1 or 2. However some population counts were not suppressed at any level. In some cases, one could still infer complementary data by subtracting data for various counties from state populations to infer population results for a county that had been suppressed.

**1990 Census** The 1990 census was the first to employ the concept of swapping. The 100% data (namely PL94, Summary File 1 and Summary File 2) were published down to the block level. But, where there was risk of potential disclosure, a confidentiality edit was performed on the census microdata. For those small blocks deemed at risk, Census selected a small sample of households with a higher sampling rate of such at-risk households used in small census blocks. These at-risk records were paired with other census records from other geographic locations using a set of matching rules. The matching process preserved key attributes such as household size, the age of those residing in a given location, etc. The household records are then swapped and the interchanged version is what is used for the Census Edited File that then forms the source of the various tabular summaries. The rate of swapping is not disclosed so as to prevent possible reverse engineering of the process. In addition, Census began using rounding of entries as well as top and bottom coding to prevent respondent identification arising from

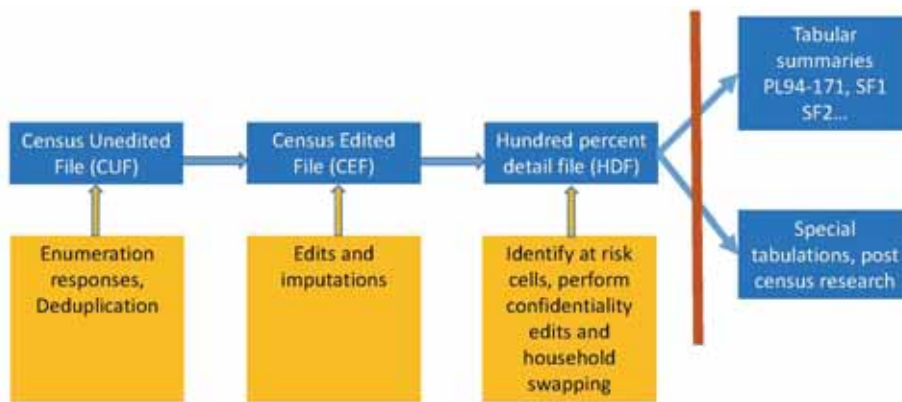


Figure 3-9: A graphical depiction of the disclosure avoidance process used in the recent 2010 census.

age extremes etc.

**2000 Census** For the 2000 census, more emphasis was given to protecting small blocks and block groups from possible re-identification. For this census, the race category was expanded to include 63 possible alone or combined races. The probability of swapping was increased to those cases where disclosure risk was thought to be higher such as cross-tabulations of key variables, smaller blocks, and also households that contained unique races in that census block.

**2010 Census** The approach to disclosure avoidance used in 2010 largely followed the approaches developed in the earlier 1990 Census as discussed above. In addition, Census developed partially synthetic data for group quarters in which it blanked values that were assessed as at risk and instead substitutes those values with data obtained from regression models. In summary the disclosure avoidance process follows steps outlined graphically in Figure 3-9. In the next section we discuss why this approach was ultimately judged inadequate.

---

## 4 THE CENSUS RE-IDENTIFICATION VULNERABILITY

In this section we discuss the vulnerability discovered by Census using the 2010 census data. We then examine the fundamental basis of the vulnerability: the results demonstrated in 2003 by Dinur and Nissim [5] that releasing an overly large number of statistics about a database allows one to perform reconstruction of the detailed data comprising that database. This result holds true even when one tries to preserve privacy by noising the results of database queries. We verify some of their observations in this section. We also offer a reinterpretation of their results in terms of information theory. Our discussion essentially validates the conclusion of Census that it is possible to reconstruct census microdata even after the application of traditional disclosure avoidance techniques like swapping, top and bottom coding etc.

### 4.1 Reconstruction of Census Tabular Data

The tabular summaries found in Census products such as PL94-171, SF1 and SF2 have been viewed in the past as safe to publish. These summaries are built using census microdata and it is this microdata that is controlled via disclosure avoidance. For the 2010 census the techniques discussed in Section 3.3 were all used; randomized swapping of households, top and bottom limitations on populations and ages, etc.

In 2018 Census looked at the feasibility that the tabular summaries could be processed to infer the microdata records that were used to produce them [1]. This had not been thought to be feasible owing to the large amount of data and computation involved. Such reconstruction of the microdata is not yet a violation of Title 13 since no personal data (e.g. names, addresses, etc.) are used when these tables are built. But, as in other re-identification attacks, if external data can be joined with the microdata then it may be possible to relink the microdata with



---

the associated personal data.

In creating the major products published by the Census, each time a cell is populated in a table it is a result of a query made on the microdata. For 2010 the number of queries (or equivalently the number of tabulations ) in the PL94 publication is about 3.6B or about 10 for every person in the US. For SF1, the number of tabulations is 22B for population and 4.5B for tabulations of households or group quarters. For SF2 there are 50B tabulations. And for the survey of American Indians and Alaskan Natives there are 75B tabulations. Thus Census publishes a total of 155B queries over the population and households of the US. The population of the US in 2010 was approximately 310M and so many more queries than people (by a significant multiple) have been issued. Most of the microdata entries used to produce these tables have not been processed through traditional disclosure methods.

To test the likelihood of reconstruction Census selected only a subset of the tables that are published. These were

P001	Total population by block,
P006	Total races tallied by block,
P007	Hispanic or Latino origin by race by block,
P009	Hispanic or Latino and not Hispanic or Latino by race by block,
P011	Hispanic or Latino and not Hispanic or Latino by race by age ( $\geq 18$ ) by block,
P012	Sex by age by block,
P012A-I	Sex by age by block iterated by race,
P014	Sex by age ( $< 20$ ) by block,
PCT012012A-N	Sex by age by tract iterated by major race alone.

Each table entry is equivalent to an integer-valued linear equation over the microdata tables. For example, if we set the count of people in tract  $t$  who are

---

male and who are 27 years old to  $T_{t,M,27}$  then this is tabulated via the equation

$$T_{t,M,27} = \sum_p \sum_r \sum_b B_{p,M,27,r,b}, \quad (4-1)$$

where  $p$  sums over the internal person number in the microdata,  $r$  sums over the possible races, and  $b$  sums over the block codes associated with tract  $t$ . The summand  $B$  is a selector that is 1 if a record indicates a male of age 27 of any race residing in a block in tract  $t$  and zero otherwise [17]. The sum over race is necessary to pick up one of the 63 combinations of race recognized by OMB.

To solve the resulting collection of equations, Census used a state of the art optimization solver known as Gurobi [12]. The Gurobi solver attempts to find the best integer solution to the set of equations corresponding to the tabulations. To break up the problem into manageable pieces Census applied the solver at the tract level. The solver was able to solve the resulting systems with few exceptions. The microdata for the entire US was determined in this way for all 70,000 Census tracts and all 11M Census blocks. To perform the relevant calculations, a virtual parallel cluster was instantiated using Amazon Elastic Cloud facilities and, for this workload and cluster configuration, completed the task in several weeks. Such a task therefore is not outside present day capabilities.

The resulting reconstructed microdata contained

- A geocode at the block level
- A binary variable indicating Hispanic origin (or not) and one of the 63 possible OMB race categories
- Sex
- Age (by year).

Census does publish a sample of the microdata called the Public Use Microdata Sample (PUMS) for use by demographers and other researchers for both the decennial census and for the American Community Survey, but these are rigorously

---

curated to make sure individual information cannot be inferred. For example, the geographic resolution is limited to areas with populations over 100000. In contrast, the reconstructed data has no population threshold and contains data like single year ages, race, and ethnicity at the block level.

The next step was to see if the reconstructed microdata could then be linked with commercially available marketing data. Some of this data is freely available or could be reconstructed using public records, but more complete and current databases can be licensed through marketing research firms. Such commercial data typically contain names, addresses, sex and birthdate but typically do not contain information regarding race and ethnicity. While not investigated in this case, Census data also contain information about family make-up. Using the reconstructed database, and acquiring commercial data, Census performed a database join using the age, sex and block locations as the common columns of the two datasets. The entries in the resulting table would now have the name and address of the respondent. If correct, these would be a re-identification of the microdata records. Release of this information would constitute a violation of Title 13.

Census determined that 46% of the reconstructed records matched correctly to the internal microdata. If a fuzzy match on age were used, 71% of the records matched. Thus the reconstruction algorithm using only some of the Census tables matched correctly 71% of the US population. Of those internal Census records, 45% were successfully mapped to a corresponding record in a commercial database again using fuzzy age matching with a one year uncertainty. Census then took the records that matched to see if they in turn matched the internal records Census collects when people submit their responses that contain name and address. Of the records that matched the commercial data sets, 39% of these matched exactly with Census records. This corresponds to the successful re-identification of 52M people or 17% of the population in 2010. Previous estimates of the re-identification rate was 0.017% of the population and only 22% of these were confirmed to be correct. The re-identification risk demonstrated by Census is four orders of magnitude larger than had been previously assessed [27].

---

In section 4.2 we examine a simplified version of this reconstruction problem in which the data set is just a column of bits to verify that the type of attack described above is not specific to the data protected by the Census. It is a general difficulty associated with publishing too many query results about a sensitive dataset.

## 4.2 Results of Dinur and Nissim

As discussed in Section 4, a key motivation for the development of formal privacy approaches to further secure the 2020 census is the Fundamental Law of Information Recovery. This observation, as quoted by Dwork is that

“overly accurate estimates of ‘too many’ statistics is blatantly nonprivate.”

By blatantly nonprivate is meant that given some database with information we wish to keep private there exists a methodology to issue queries on the dataset that will allow one to infer a dataset whose elements differ from the original in some number of elements. The number of elements that are not obtained correctly reduces as the size of the database increases. Thus for a large enough database the methodology asymptotically extracts all the elements of the private database.

Dinur and Nissim [5] demonstrated this in a seminal paper by modeling a database as a set of binary numbers whose (private) values we are interested in learning. The database is represented by an array of binary digits:

$$d = (d_1, d_2, \dots, d_n). \quad (4-2)$$

A *statistical query* is represented by a subset  $q \in [1, 2, \dots, n]$ . The exact answer to the query is the sum of all the database entries specified by  $q$ :

$$a_q = \sum_{i \in q} d_i. \quad (4-3)$$

---

An answer  $A(q)$  is said to be within  $\epsilon$  perturbation if

$$|a_q - A(q)| \leq \epsilon$$

The *algorithm*  $A$  is said to be within  $\epsilon$  perturbation if for all the queries  $q \subseteq [n]$  the answers  $A$  are within  $\epsilon$  perturbation. Dinur and Nissim define the notion of  $T(n)$  non-privacy if there exists a Turing machine that terminates in  $T(n)$  steps so that the probability of determining any fraction of the bits with the exception of a vanishingly small number as the size of the data set increases is essentially one. The result of most relevance to this study is that if the query algorithm provides  $o(\sqrt{n})$  perturbation then non-privacy can be achieved with an algorithm that terminates in a number of steps that grows polynomially with increasing data set size. More noise than this is required to get even weak privacy. Dinur and Nissim describe an algorithm using linear programming to demonstrate the existence of such an algorithm. The conclusion is that, even in the presence of noise, a sufficiently capable adversary can infer the secret bits of the dataset. In order to ensure privacy one must restrict the number of queries or add so much noise that the utility of statistical queries on the dataset is potentially degraded.

### 4.3 JASON Verification of the Dinur-Nissim Results

JASON undertook a verification of the Dinur-Nissim results using a variation of their approach. First we examine the situation where no noise is added to the queries. We then examine the situation where we add noise. We begin by generating a random vector of zeros and ones,  $\mathbf{d}$ , of size  $n$ . We then create an  $m \times n$  random matrix,  $Q$  of zeros and ones. These will be the queries. We then compute the matrix vector product of the query matrix with the database vector. These are the random query results. We then use bounded least squares with constraints to solve the following problem:

$$\operatorname{argmin} \|\mathbf{Q}\mathbf{x} - \mathbf{d}\|^2 \text{ subject to } 0 \leq x_i \leq 1. \quad (4-4)$$

Once this problem is solved we then round the components of the resulting vector

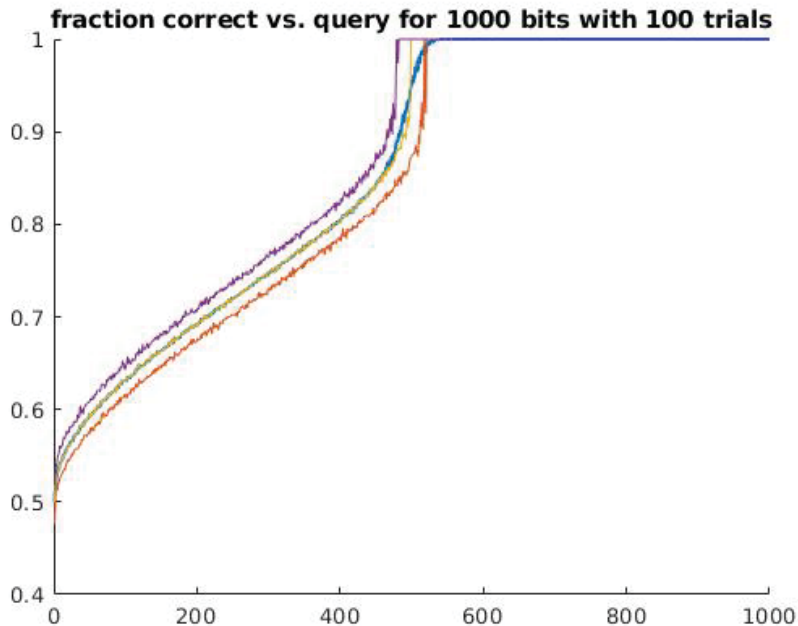


Figure 4-1: Fraction of bits recovered for a 1000 bit Dinur-Nissim dataset as a function of the number of random queries. The lower curve is the minimum fraction recovered, the middle curve is the mean, and the upper curve is the maximum recovered. No noise is added to the query results.

$x$  to 0 or 1. If we issue  $n$  queries and our query matrix is not singular,<sup>2</sup> then we would recover the results of the database immediately. But in fact the full database can be recovered with less than  $n$  queries in the absence of noise. In Figure 4-2 we plot the fraction of bits computed correctly as a function of the number of queries for a database of size 1000 bits. Because our queries are random we perform 100 trials and plot the 10% decile of the fraction of bits recovered (lower curve), the 90% decile fraction of bits recovered (upper curve) and the mean recovered (middle curve).

With no queries we recover 50% of the bits, but this is of course no better than random guessing. As the number of queries increases we recover more of the bits (although the bits recovered will differ with each random attempt). It is to be expected that we would recover all the bits once we issue 1000 random

<sup>2</sup>singularity would be a very rare event

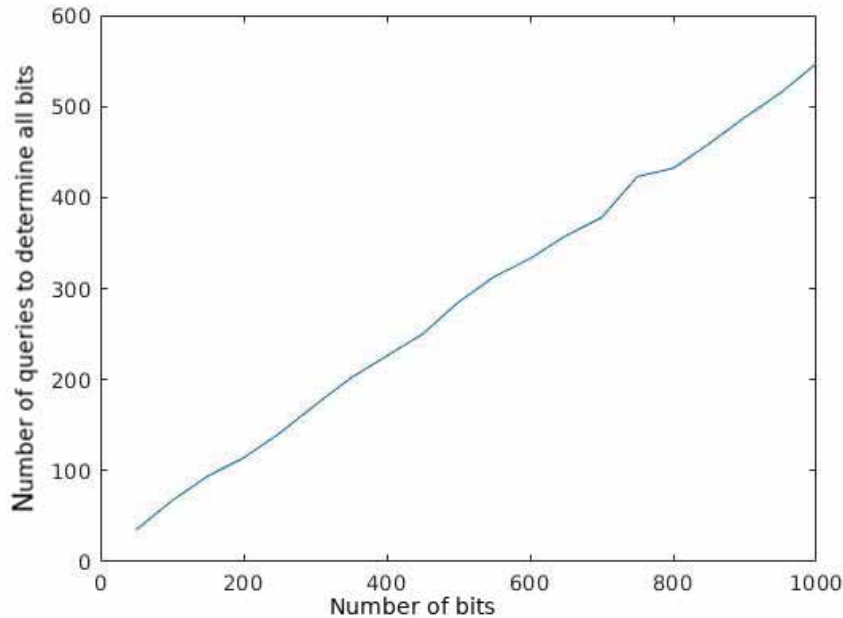


Figure 4-2: Number of queries needed to recover 100% of the private bits in the Dinur-Nissim dataset as a function of the size in bits of the data set.

queries but as is seen in the Figure all the bits are recovered at about the half way mark in the number of queries. If one repeats this calculation for databases of varying size  $n$  and asks how the number of queries required to achieve perfect knowledge of the bits varies with  $n$  one gets a roughly linear variation in  $n$  as shown in Figure 4-2. The slope of this roughly linear variation as a function of increasing database size is shown in Figure 4-3. As can be seen the slope is close to  $1/2$  indicating that roughly  $n/2$  queries are required on average to determine the entire database. This is a special aspect of this particular type of database. A random query response will get information about a number of the bits. For example, if we choose to query two bits at a time by summing the values, then a sum of zero immediately tells us the two bits must be zero. Similarly if we get a sum of 2 we know immediately the two bits we queried must have both been one. Thus one can infer the bits more quickly in a probabilistic sense than simply asking for one bit at a time which would correspond to the query matrix being the identity. In section 4.5 we apply an information-theoretic argument to show that

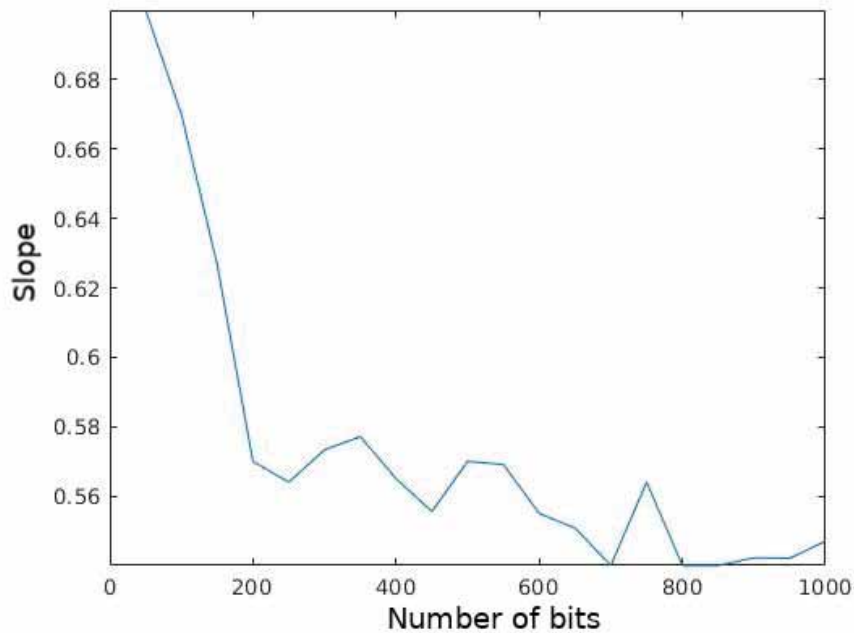


Figure 4-3: Same as Figure 4-2 but each point is normalized by the number of queries. As the number of bits increases the curve appears to approach a limit of 1/2

the results we get from our least squares approach are not far from optimal.

The results above certainly confirm that, without noise, it is possible through a sequence of queries to infer the entries of a database. It should also be noted that a recovery approach based on optimization will also succeed if one poses more queries than the number of entries in the database. To be sure, the Dinur-Nissim database is special, but it is easily confirmed that through publication of tabular summaries that comprise (sometimes multiple times) the information contained in the database, recovery of the bits, in this case a stand-in for microdata, is possible.

If we think of census data as a (very large) Dinur-Nissim database we can see that the reconstruction attack is quite plausible. In terms of bits, a rough count of the number of bits contained in the Census Edited File might be

- 3 bits to describe the 8 types of group quarters (8 levels),



- 
- 5 bits to describe a person’s age (here we assume ages are only reported in intervals of 5)
  - 1 bit to describe Hispanic origin (2 levels),
  - 6 bits to describe race (63 OMB race designations),
  - 24 bits to describe the 11 million census blocks,

for a total of 39 bits per person. If we estimate that in 2010 there were  $3 \times 10^8$  residents in the US this totals to  $1.2 \times 10^{10}$  bits. If we examine the number of queries in a full cross table this would be

$$(8 \times 20 \times 2 \times 63) \times 1.1 \times 10^7 = 2.2 \times 10^{11}$$

This rough estimate indicates that the census tables “overquery” the data set by a factor of almost 20. If we treat the Census database reconstruction effort as an attempt to infer the bits in a large Dinur-Nissim database there is no question the database (up to the edits that are used to create the tables) could be reproduced with perfect accuracy. A similar argument using the idea of Boolean satisfiability (SAT) solvers is given in [10].

#### 4.4 Queries in the Presence of Noise

Given the vulnerability discussed above it is perhaps of more interest to examine the number of queries that must be issued to recover the database when each query is perturbed by noise. To examine this, we used the same bounded least squares optimization approach but in the presence of noise. For a dataset size of  $n$  bits we added to each random sum a perturbation sampled from a normal distribution of mean 0 and variance  $\sqrt{n} \log(n)/2$  where  $n$  is again the number of secret bits in the database. The reason for this particular choice was to see if the optimization approach would fail with an increasing number of queries. According to Dinur and Nissim if one adds noise with an amplitude of greater than  $O(\sqrt{n})$  then recovery

---

should be impossible. We were unable to confirm this observation. Instead, as the number of queries increases, an increasing fraction of the correct bits is returned. This is most likely not in conflict with the theorems of Dinur and Nissim as they require that the adversary be time bounded whereas in our approach we do not impose any time limit but instead continually issue queries. The results are shown in Figure 4-4. In the Figure we show the fraction of bits determined correctly as a function of the number of queries for databases of varying size. For each database of size  $n$  we added a random perturbation sampled from a normal distribution of mean 0 and variance  $\sqrt{n} \log n / 2$  to each query.

We perform a query of size  $m$  100 times and provide some statistics for the results. The red, yellow and purple lines indicate the 10%, 50% and 90% deciles respectively of fraction of bits recovered correctly; the blue lines indicate the mean of the fraction of bits recovered correctly. As can be seen, the number of queries required increases greatly, but, in all cases, all metrics measuring the fraction of bits recovered correctly increase towards one. Thus if one is willing to issue a large number of queries, for example, a large multiple of the number of bits, eventually one will learn the internal records of the database. Apparently, the use of random queries will provide results that average out the applied noise and recover the required information. In some ways this is to be expected. For example if we were allowed to issue directly a query for bit  $i$  of the  $n$  bits in the presence of noise, we would have received a random response, but continual averaging over the responses would have recovered the result regardless of the amount of noise. Indeed we would have predicted that we would have required a number of queries which is some constant factor of the variance. We discuss this further in Section 5 where we consider how many queries are required for a given noise level to recover the internal bits. In the next section we apply information theory to compute idealized estimates of the number of queries required to infer the internal data of the Dinur-Nissim database both in the absence and presence of noise.

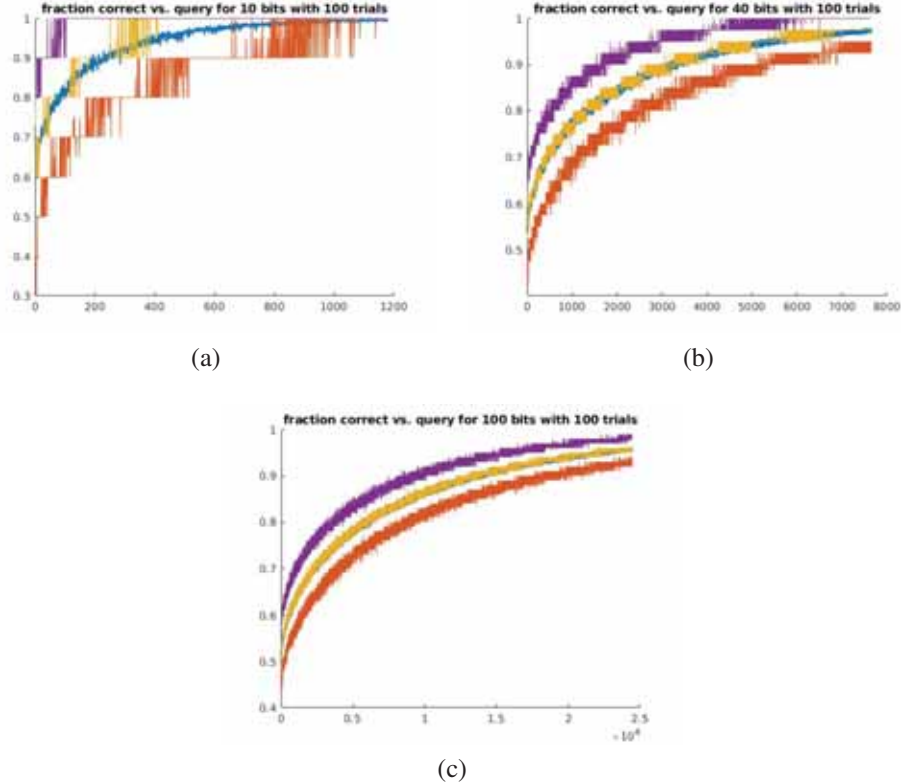


Figure 4-4: Fraction of bits recovered as a number of queries for databases of size 10, 40, and 100 bits. For each case we infuse the query results with Gaussian noise of means 0 and variance  $\sqrt{n} \log n/2$ . The red, yellow and purple lines indicate the 10%, 50% and 90% deciles respectively of fraction of bits recovered correctly; the blue lines indicate the mean of fraction of bits recovered correctly. Note that as the number of queries increase, the fraction of bits recovered grows until all the bits are recovered with near certain probability.

## 4.5 Information Theory and Database Uniqueness

The purpose of this subsection is to look at Dinur & Nissim's [5] fundamental results about database reconstruction from alternative points of view, namely linear algebra and (especially) information theory. The discussion is rather lengthy (but we hope pedagogical) so we have relegated it to an Appendix, but we summarize the main results:

1. In the absence of noise, a database of  $n \gg 1$  bits is determined by the re-

---

sults of approximately  $2n/\log_2 n$  queries, on the average over all possible databases. Put differently, we can expect to recover most of the bits of most databases.

2. If noise with variance  $\sigma_N^2 < n/48$  is added to the results of each query, the database remains determined by no more than  $\sim n$  queries on average.
3. If the noise variance  $\sigma_N^2 \gg n/16$ , we expect to require  $\sim 16\sigma_N^2$  queries to fix the bits uniquely.

It should be noted that there are at least two facets to DN's results: (i)  $o(\sqrt{n})$  noise allows the database to be uniquely specified using algebraically (in  $n$ ) many queries; and (ii) the bits can actually be reconstructed in polynomial time using linear programming. Apart from a few obvious remarks about linear algebra in the noiseless case, we have nothing to say here about the computations required to do the actual reconstruction. Our information-theoretic arguments advanced here are nonconstructive, in much the same way as the Shannon channel-capacity theorem [31], which does not say by what encodings the capacity can be achieved.



---

## 5 DIFFERENTIAL PRIVACY

The Census has proposed the use of Differential Privacy (DP) as the basis for its future Disclosure Avoidance System (DAS). The goal of DP is to prevent one from learning about the possible participation of an individual in a survey. The idea is that the result of a query into the dataset provides results that are largely the same even if an individual opted out of participating in the survey. This is accomplished by adding noise to the results of queries so that one cannot easily perform the types of record linkage attacks that have determined the details of database records from queries in the past. DP introduced by Cynthia Dwork [7, 8] and colleagues and developed since then in a vast research literature is viewed as the present gold standard for formal privacy guarantees. The definition is phrased in a language that may be unfamiliar, so we go over it in detail.

The setting is databases and database queries. A database  $D$  is a collection of records. Each record has attributes (age, sex, HIV-positive, wealth, or whatever), and each attribute has a range of values it can take. A query is just some function on the database. For instance, “how many records are there”, “what is the average age of HIV-positive people”, and so forth. We think of attributes being exact and queries giving precise answers, but that is not always desirable as we have discussed previously and is in fact a mental shortcut. Age is reported in years, not days, so people with age 12 are those aged between 12 and 13. Then average age is also reported in years, not some exact number like 62381/129.

DP is a property of algorithms for answering queries. It is clear that, to preserve privacy, queries cannot just return the right answer, so one can think of an algorithm that answers a query as adding noise to the correct answer. Adding noise means that the algorithm is not deterministic, but probabilistic, using random numbers. The approach in which noise is added to the query is known as a mechanism.

---

An algorithm  $\mathcal{A}$  is  $\epsilon$ -DP ( $\epsilon$ -differentially private) if

$$e^{-\epsilon} < \frac{\Pr(\mathcal{A}(D) \in T)}{\Pr(\mathcal{A}(D') \in T)} < e^{\epsilon}$$

where  $D$  and  $D'$  are any two databases that differ by one record. The probabilities come from the random numbers that  $\mathcal{A}$  uses.  $T$  is the set of possible outcomes of  $\mathcal{A}$ . For instance, if the query was for average age, then  $T$  would be an interval like  $[37,38)$ , meaning that the average age is between 37 and 38. Alternately, if  $\mathcal{A}$  returns continuous values, then one needs to measure the probability that the result lies in an interval, rather than takes on a specific value.

A key element of DP is the notion of the privacy budget. In the DP literature this is typically labeled  $\epsilon$ . The notation is set up so that a value of  $\epsilon = 0$  indicates zero privacy loss. The technical definition of a DP algorithm is as follows:

**Theorem.** *An algorithm  $\mathcal{A}$  satisfies differential privacy if and only if for any two datasets  $D$  and  $D'$  that differ in only one record, we have that for all results  $T$  that lie in the range of the algorithm  $\mathcal{A}$*

$$\Pr[\mathcal{A}(D) \in T] \leq \exp(\epsilon) \Pr[\mathcal{A}(D') \in T].$$

*Equivalently the ratio of probabilities*

$$\frac{\Pr[\mathcal{A}(D) \in T]}{\Pr[\mathcal{A}(D') \in T]} \leq \exp(\epsilon).$$

Note that there is nothing special about  $D$  and  $D'$  so we can write the inequality in a symmetric two-sided manner as we did above:

$$\exp(-\epsilon) \leq \frac{\Pr[\mathcal{A}(D) \in T]}{\Pr[\mathcal{A}(D') \in T]} \leq \exp(\epsilon).$$

If an algorithm satisfies the definition of being differentially private, the expression above provides a bound on how much additional information one can infer from adding or deleting a record in a database. This will prevent learning about a specific record through the examination of the two datasets for example through database differencing. It also makes record linkage attacks more difficult in that it introduces uncertainty in the query results.

---

Perhaps of more importance, DP algorithms by definition provide formal bounds on how many queries can be made before the probability of learning something specific about a database increases to an unacceptable level. This is the real role of the privacy budget. A DP algorithm with a large value of  $\epsilon$  indicates that the ratio of probabilities of learning a specific result in two datasets with one record differing is large and so implying that the query using the algorithm discriminates strongly between the two datasets. On the other hand, a small value of  $\epsilon$  means little additional information regarding the dataset is learned. It is not hard to show that DP has several properties that make it possible to reason about how the privacy budget is affected by queries.

**Sequential access to the private data degrades privacy** Suppose we have an algorithm  $\mathcal{A}_1$  that satisfies DP with privacy loss parameter  $\epsilon_1$  and another algorithm  $\mathcal{A}_2$  that has a privacy loss parameter  $\epsilon_2$ . If both algorithms are composed then the privacy loss parameter for the composed algorithm is the sum of the individual privacy loss parameters. we have

$$\begin{aligned} \Pr[\mathcal{A}_2(\mathcal{A}_1(D), D) = t] &= \sum_{s \in \mathcal{S}} \Pr[\mathcal{A}_1(D) = s] \Pr[\mathcal{A}_2(s, D) = t] \\ &\leq \sum_{s \in \mathcal{S}} \exp(\epsilon_1) \Pr[\mathcal{A}_1(D') = s] \exp(\epsilon_2) \Pr[\mathcal{A}_2(s, D') = t] \\ &\leq \exp(\epsilon_1 + \epsilon_2) \Pr[\mathcal{A}_2(\mathcal{A}_1(D') D') = t]. \end{aligned}$$

In general, if one composes this way  $k$  times the effective  $\epsilon$  becomes

$$\epsilon = \epsilon_1 + \epsilon_2 + \dots + \epsilon_k.$$

This implies that one must account for all the operations to be performed on the data in order to ensure a global level of privacy over the whole dataset. It also demonstrates, at least in terms of bounds, the cost of a number of queries on a database in terms of overall privacy and that repeated queries on the data will boost the ratio of probabilities. This provides a useful quantitative aspect to assessing disclosure risk although it is not explicitly a statement about disclosure risk.



---

**The privacy budget behaves gracefully under post-processing** If an algorithm  $\mathcal{A}_1$  satisfies DP with a privacy budget of  $\epsilon$ , then for any other algorithm  $\mathcal{A}_2$  which post-processes the data generated by  $\mathcal{A}_1$ , the composition of  $\mathcal{A}_2$  with  $\mathcal{A}_1$  satisfies DP with the same privacy budget. To see this, suppose  $S$  is the range of the algorithm  $\mathcal{A}_1$ . Then we have

$$\begin{aligned} \Pr[\mathcal{A}_2(\mathcal{A}_1(D)) = t] &= \sum_{s \in S} \Pr[\mathcal{A}_1(D) = s] \Pr[\mathcal{A}_2(s) = t] \\ &\leq \sum_{s \in S} \exp(\epsilon) \Pr[\mathcal{A}_1(D') = s] \Pr[\mathcal{A}_2(s) = t] \\ &\leq \exp(\epsilon) \Pr[\mathcal{A}_2(\mathcal{A}_1(D')) = t]. \end{aligned}$$

It is important in this argument that only the algorithm  $\mathcal{A}_1$  accesses the private data of the database. This composition property is quite powerful. One of its most important applications is that if you transform the database into another database with synthetic data processed through a DP algorithm then additional processing of that data will preserve differential privacy. Thus one can create a dataset from the original dataset and preserve differential privacy for future processing of the synthetic data. This feature is an important component of the disclosure avoidance system currently under consideration by Census.

**Parallel composition** If one deterministically partitions a database into separate parts then one can control the privacy loss. If  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_k$  are algorithms that respectively only access the (nonoverlapping) partitions of the database  $D_1, D_2, \dots, D_k$  then publishing the results of the queries  $\mathcal{A}_1(D_1), \mathcal{A}_2(D_2), \dots, \mathcal{A}_k(D_k)$  will satisfy DP but with an  $\epsilon$  given by

$$\epsilon = \max(\epsilon_1, \epsilon_2, \epsilon_k).$$

Such results show that the production of a histogram where the data is partitioned into categories and then counts are published for each category can still preserve a given privacy budget.

---

## 5.1 Mechanisms

The definition of DP does not guarantee that there are any DP algorithms, but of course there are. In general, a *mechanism* is a way of generating DP algorithms from data base queries. We discuss some of these below.

### 5.1.1 Laplace mechanism

Consider a query whose correct answer is some continuous numeric value. The query has sensitivity  $\Delta$  if the correct answer on any two neighboring databases  $D, D'$  can differ by at most  $\Delta$ . Then an  $\epsilon$ -DP algorithm for this query would add  $\text{Lap}(\Delta/\epsilon)$  noise sampled from a Laplace probability distribution to the correct answer, where  $\text{Lap}$  is the two-sided Laplace distribution. The probability density for the Laplace distribution with parameter  $\beta$  is

$$\frac{1}{2\beta} \exp(-|x|/\beta).$$

More usefully, to generate a random Laplace variate from a uniformly distributed  $p$  between 0 and 1, one can compute

$$\beta \operatorname{sgn}(p - 0.5) \ln(1 - 2|p - 0.5|).$$

This density has mean 0 and a variance of  $2\beta^2$  and is displayed in Figure 5-1. In applications to DP we use the relation  $\beta = 1/\epsilon$ . Thus small values of privacy loss imply large values of  $\beta$  and so very broad distributions with large variances. Note that the use of the Laplace mechanism and the associated Laplace distribution matches exactly with the definitions of DP in terms of the bounds on probabilities. Other distributions can be used, for example, a normal distribution, but in this case there may be small violations of the DP bounds for extreme values. A slightly modified definition of DP is required to handle this case but its use would not affect our conclusions so we won't discuss it further.

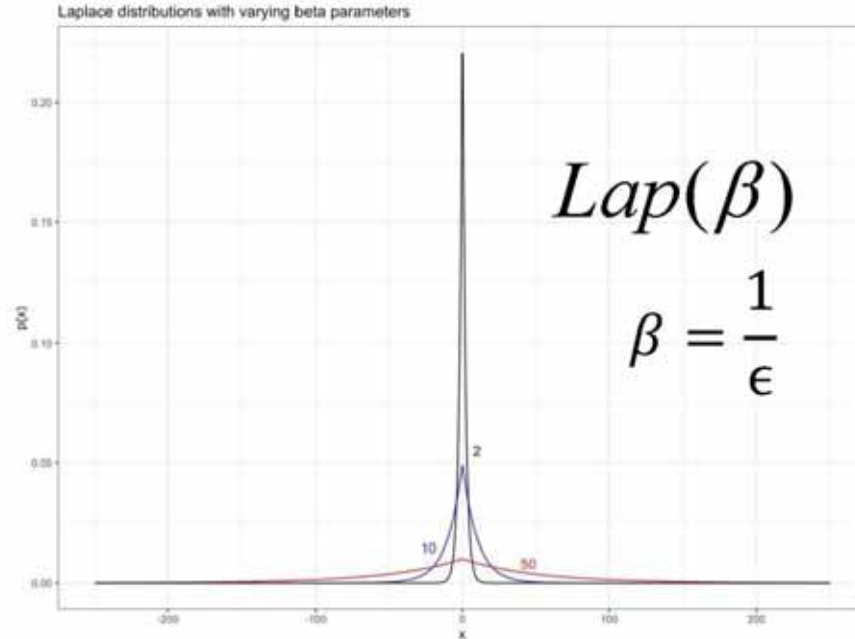


Figure 5-1: The Laplace distribution for several values of the parameter  $\beta$ . A large  $\beta$  corresponds to broad tails.

### 5.1.2 Geometric mechanism

The Laplace mechanism does not produce integers for integer-valued attributes. The Geometric mechanism adds an integer to the correct answer, where the integer is randomly chosen from a suitable geometric distribution. One could instead use the Laplace mechanism and round, but these results are slightly different. The (two-sided) geometric distribution with parameter  $\alpha$  has probability density

$$\frac{\alpha - 1}{\alpha + 1} \alpha^{-|x|}$$

for producing integer  $x$ . If  $\Delta$  is the sensitivity of the query,  $\epsilon$ -DP is the same as  $\alpha = \exp(\epsilon/\Delta)$ .

---

### 5.1.3 Matrix mechanism

In applying DP to the census tables one approach would be to make one colossal query of the confidential data that produces at once all the tables that the public will be able to see. Each number in each of these tables is a count, so the colossal query can be represented as a big matrix  $M$  applied to a huge vector  $c$  of the confidential data. DP would add noise to each count in  $Mc$ . But this may introduce more noise than is strictly required. A way to deal with this is known as the matrix mechanism [25, 19]. The public tables published by the Census are counts over discrete categories. The (confidential) data is a data base where each record has some attributes, and each attribute only takes on a finite set of values. These include age (from 0 to some upper bound), sex, Hispanicity, race (63 values), and so forth. An equivalent way of representing the data is as a (long) histogram, with one count for each possible combination of attributes. So there would be a count for ‘male black-asian hispanics of age 37’ and one for ‘female white non-hispanics of age 12’, and so forth. If these are arranged in some arbitrary order, we can think of the data base as a vector of counts  $(x_1, x_2, \dots, x_n)$ . Then the result of a count query (e.g., ‘male native-americans’) is the inner product  $w \cdot x$  where  $w$  is a vector of 0s and 1s of length  $n$ , with 1s exactly for those places in the histogram that count male native-Americans. This inner product is one of the counts in the publicly released tables. The set of queries that produce all these counts can be represented as the rows of a very large matrix  $W$ .

The idea of the matrix method is to answer all these queries (or this one giant query) in two stages. First answer a set of *strategy* queries in a differentially private way, and then combine the answers to these queries to get the queries we want ( $Wx$ ). The strategy queries can be represented by some matrix  $A$ , one computes  $m = Ax + \Lambda$ , where  $\Lambda$  is a vector of noise chosen so that the result is  $\epsilon$ -DP. Then any post-processing of  $m$  does not affect privacy, so if  $W = UA$ , then  $Wx = Um$ , which are the tables we want. One can attempt to find such an  $A$  that minimizes the mean error in the output. The process is illustrated graphically in

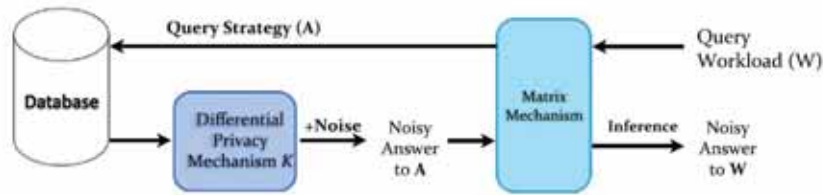


Figure 5-2: Process utilized by the matrix mechanism (from [25]).

Figure 5-2. This is a substantial computation described in the referenced papers.

## 5.2 Some Surprising Results in Applying Differential Privacy

The definition of DP does not immediately speak to the kinds of errors introduced. Nor does it guarantee that a query has a satisfactory (or any) DP algorithm. Below are presented some examples that indicate that one must be careful sometimes with the result of DP calculations to ensure statistical utility of the results.

### 5.2.1 Cumulative distribution functions

In [26] an example is given of how DP can affect common statistical measures. For example if we want to compute a cumulative distribution function (CDF) of incomes in some region we would count the number of income values less than some prescribed value and then divide by the total number of incomes to get a distribution. Under DP each time such a query is issued noise is added to the result. Depending on the level of noise injected the resulting CDF may become non-monotonic, something that is mathematically forbidden. Some results are shown in Figure 5-3 for a sample CFD under various values of  $\epsilon$ . As  $\epsilon$  is increased the generated CFD will converge to the smooth case without noise. The examples shown with a large amount of injected noise could not for example be reliably differenced to provide probabilities over small intervals. This is in fact the point - we cannot focus too clearly on the small scales. The issue identified here can be easily fixed by re-sorting the data so that a monotonic CFD results. The main

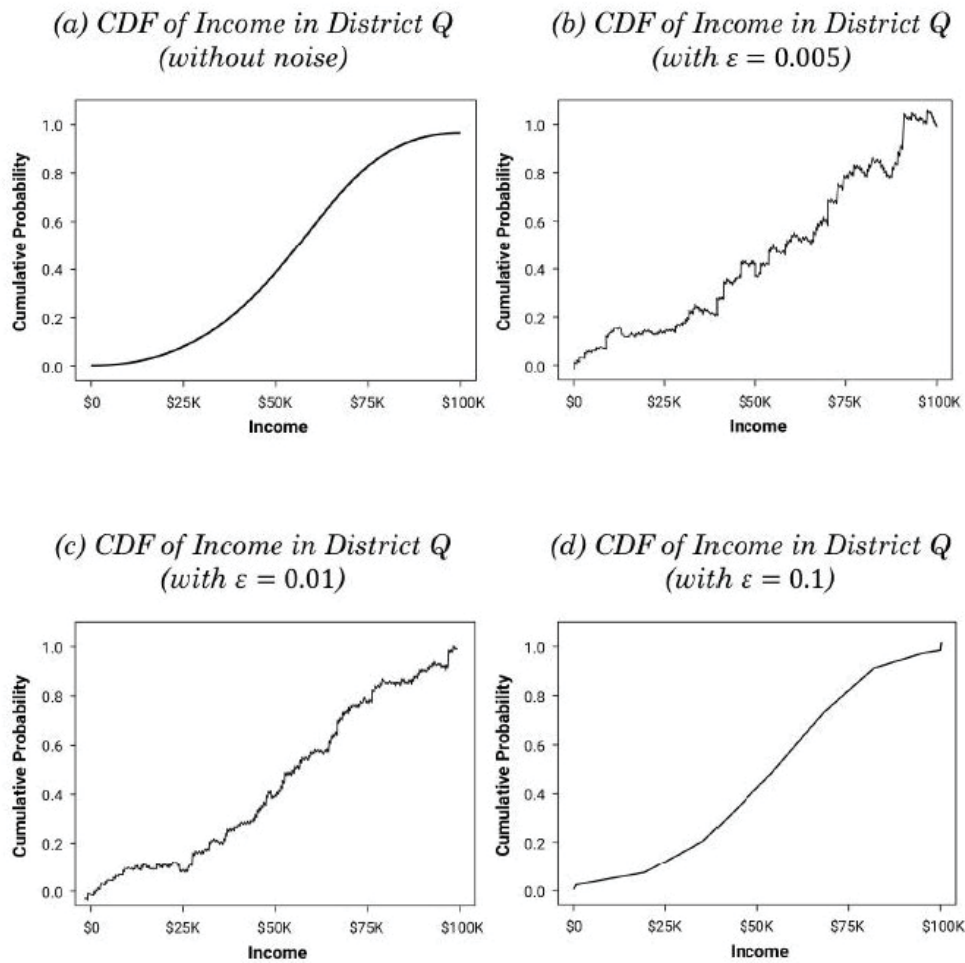


Figure 5-3: An example of a CDF of incomes under various values of the privacy loss parameter (from [26]).

point here is simply to point out possible issues with results published directly under DP.

## 5.2.2 Median

The examples of mechanisms so far involve additive noise, but the definition does not mention the type of noise. Consider a query that asks for the median. If the middle three elements in the larger database are 0.12, 0.14, 0.19, then if the size

---

of the database is odd, the median is 0.14, otherwise some tie-breaking algorithm would be needed. The smaller database is the result of removing one record from the larger database. If the number removed is no more than 0.12, the new exact median will be between 0.12 and 0.19. If the number removed is 0.19 or more, the same is true, and if 0.14 is removed, it is also true. So a privacy algorithm could choose any number between 0.12 and 0.19. Note that this algorithm decides what to do based on the data. It satisfies the intuition behind DP in that the result is independent of which record is removed from the database. However, it is *not*  $\epsilon$ -DP for any  $\epsilon$ . To see that, consider what the algorithm returns for the smaller database, if 0.12 were returned. Then the middle 3 might be 0.10, 0.14, 0.19, and the algorithm could return any value between 0.10 and 0.19. In particular there is a positive probability of returning a value in the interval  $[0.10, 0.12]$  for the smaller database, but that's impossible for the larger. So the ratio of probabilities in the definition of DP would be 0, which is impossible for any  $\epsilon$ .

For the median, however, the sensitivity  $\Delta$  is large. If the attribute takes on values between 0 and 1, and in the smaller database half of them are 0 and half of them are 1, then the median for the larger database is whatever value was removed, so  $\Delta = 1/2$  (assuming that the algorithm chooses the midpoint for even sized databases). The Laplace mechanism doesn't look at the data, so it will add  $\text{Lap}(1/2\epsilon)$  noise. Answers that then fall outside  $[0,1]$  presumably would be truncated to be in range, so there is a positive probability of getting 0 or 1, which will almost always be silly and completely uninformative.

There is a similar story for any quantile, or the min, or the max, but the median is often used as a robust measure of location. Dwork and Lei [6] give a different algorithm that should be generally more satisfactory, but is data-dependent, and can fail (returning  $\perp$  (null) in the language of computer science) on weird databases, such as the one in this example.

The decennial census data is just counts, so the peculiarities of medians are not directly relevant, but other statistical agencies and other statistical products

---

might not be so lucky.

### 5.2.3 Common mechanisms can give strange results for small $n$

Another mechanism is known as the random or uniform mechanism (UM). For a query that has a finite range, the random mechanism just chooses one uniformly; For example for the range of integers 0 through 10, choose a query response with probability  $1/11$ . The random mechanism is  $\epsilon$ -DP for any  $\epsilon$ . If one were to propose a mechanism for a query associated with this finite collection of integers, it would seem undesirable for it to give the correct answer less frequently than the random mechanism does. That is, there may be many DP algorithms for the query, and it is unsatisfactory to choose one whose accuracy (meaning the chance of getting the right answer) is less than just choosing a result at random. For small  $n$ , both the truncated Laplace or Geometric mechanisms are unsatisfactory in this way.

There are various mechanisms for producing DP count data, The simplest way to think about these is to assume the data base has records with one sensitive field that has value 0 or 1. Suppose the query that counts the number of 1s needs to be protected. We know the answer is in the range  $[0, n]$ , so the mechanism needs to produce a value in that range. The Range Restricted Geometric Mechanism (GM) produces

$$\min(n, \max(0, a + \delta))$$

where  $a$  is the true answer and  $\delta$  is an integer chosen (at random) from a geometric distribution

$$(1 - \alpha)^{|\delta|} / (1 + \alpha)$$

where  $\alpha = \exp(-\epsilon)$  and  $\epsilon$  is the parameter in differential privacy. Unfortunately, in this case, 0 and  $n$  will be over-represented. Worse, for most probability distributions on  $a$ , the actual count, if  $n$  is 2, the true answer of 1 is less likely than either of the incorrect answers 0 or 2. This is clearly a small  $n$  phenomenon,



---

but for small and modest-size  $n$  the usual mechanisms with various common loss functions give counter-intuitive results (cf. e.g. [4]).

Any mechanism for this problem is characterized by a (column) stochastic matrix  $P$ , where  $P_{i,j}$  is  $\Pr(i|j)$ , the probability the mechanism returns  $i$  when the true result is  $j$ .  $P$  is an  $(n+1) \times (n+1)$  matrix. The uniform or random mechanism (UM) has  $P_{ij} = 1/(n+1)$ , that is, choose any answer at random. The set of all mechanisms can be defined by linear equations and inequalities. The only unobvious one, differential privacy, is expressed by

$$P_{i,j} \geq \alpha P_{i,j+1}, \quad P_{i,j+1} \geq \alpha P_{i,j}$$

for all  $i$  and  $j$ . The choice of a mechanism then comes down to minimizing some loss function over this polytope, preferably by linear programming. There are  $n^2$  variables and a quadratic number of constraints.

Cormode's paper [4] notes that one can add a number of intuitively desirable constraints on the mechanism by adding linear constraints to this formulation. For instance, one might like the probability the mechanism returns the correct answer to be at least as large as the chance UM returns it,  $P_{i,i} \geq 1/(n+1)$ . Interchanging the values 0 and 1 in the statement of the problem converts a true answer  $a$  into  $n-a$ . One would expect the mechanism to be oblivious to this choice, which imposes a symmetry constraint  $P_{i,j} = P_{n-i,n-j}$ . One would like the correct answer to be at least as probable as any other. The geometric mechanism (GM) satisfies these only for sufficiently large  $n$ , at least  $2\alpha/(1-\alpha)$ , which is roughly  $2/\epsilon$ . If one adds the condition that answers closer to the true answer should be more likely than answers further away, then GM requires  $\alpha < 1/2$ .

For completeness, here is the explicitly fair mechanism of [4], which looks more complicated than it is, and satisfies their various sensible conditions:

$$P_{i,j} = \begin{cases} y\alpha^{|i-j|}, & \text{if } |i-j| < \min(j,n-j) \\ y\alpha^{\lceil \frac{|i-j| + \min(j,n-j)}{2} \rceil} & \text{otherwise} \end{cases}$$

---

where

$$y = \frac{1 - \alpha}{1 + \alpha - 2\alpha^{n/2+1}},$$

so the probability of returning the correct answer is a little larger than in the geometric mechanism, and the probabilities drop off more slowly with distance from the correct answer. The paper gives rules for choosing between this mechanism and GM.

#### 5.2.4 Nearly equivalent queries with vastly different results

Suppose we have a database for which HIV-status is an attribute, with the values 0 or 1. The query might be “are more than half of the records 1?” One sensible way of answering this question using counts would be to ask for the size of the database  $n$ , and the number of ones,  $x$ , and look at the result. The returned values would have Laplacian or Geometric noise added to them, but unless the number of ones is very near 50%, the answer to the original question just pops out. A different computation, equivalent if exact results are returned, would be to ask if the median value of HIV-status is 0 or 1. As we have seen there is a positive chance of getting a meaningless answer regardless of how different the counts of zeros and ones. A more sensible query would be to ask for the average. The average is not a count query, but it has sensitivity  $1/n$  for values between 0 and 1. So a DP query would answer with  $\text{Lap}(1/n\epsilon)$  noise added to the exact answer. This error drops rapidly with increasing  $n$ .

### 5.3 Invariants

The main promise of DP is to limit the knowledge that can be gained by adding or subtracting a record from a database. Informally if we make a small change in the input data the result of the output also undergoes a small change. That this is not always the case has been shown repeatedly through linkage attacks and database differencing. However, if certain results in a database must be openly

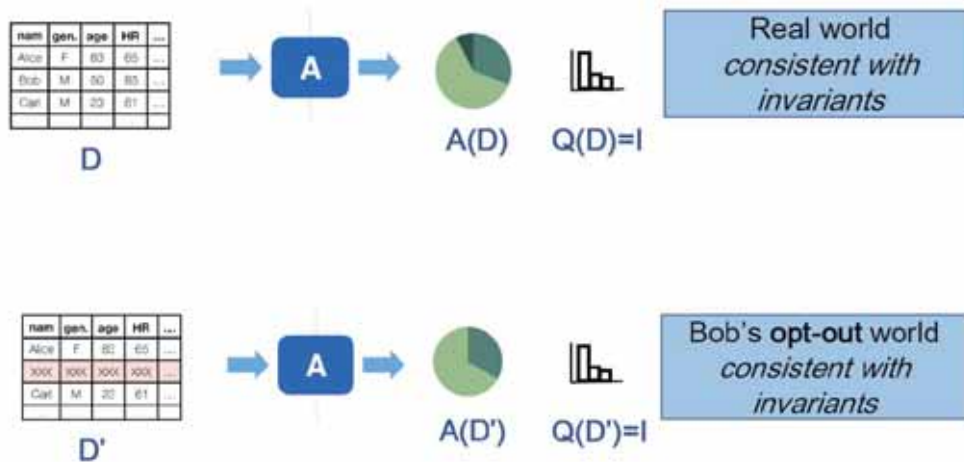


Figure 5-4: DP with invariants must be interpreted relative to a world in which respondents opt-out but consistent with invariants [21].

published without any protection then a small change in the input can have large consequence on the output if the output is directly tied to the small change.

An important example is the notion of an invariant. A simple example of an invariant relevant to the census is the need to publish an accurate count of the population of each state. For the 2020 Census, as in previous censuses, there are plans to publish state populations as exactly as possible and certainly without noise and so the state populations are invariants. In theory, releasing a true count is technically a complete violation of the DP guarantee. This is simply because removing one entry changes the population and so it is immediately obvious that a record has been removed even though we may not know which record.

As briefed to JASON by Prof. A Machanavajhala [21], it is possible to construct various scenarios where releasing an invariant could allow one to infer additional protected information regarding a record. There is to date no worst case characterization of privacy loss in this situation. At best, one can consider the incremental loss in releasing DP results in the presence of invariants. The situation is shown graphically in Figure 5-4. At present, it is not clear to what extent the

---

addition of invariants constitutes a vulnerability for Census data. As will be discussed below there are many more constraints that lead to invariants than just the population of the states. JASON does not know of a systematic approach to assess this except to perform a risk assessment by attempting to identify DP microdata as was originally performed by Census in first identifying the existing vulnerability in the absence of noise. We discuss this further in Section 7.

## 5.4 Database Joins under Differential Privacy

In creating the various Census products such as SF1, the tables are produced through a join between two databases. One contains information about persons and the other about households. Queries such as the number of men living in a particular Census block requires only access to the person database while queries such as the number of occupied houses in a Census block requires only access to the household database. But if one wants to know how many children live in houses headed by a single man this requires a join of the two databases. Joins under DP can be problematic because one must examine the full consequences of removing a record in one table as it is linked to potentially multiple records in other tables. One way to address this is to create synthetic data as the Census is doing for both tables and then perform the join as usual. This however has been shown to produce high error in the results of queries essentially because too much noise is added for DP protection. A number of groups have researched this issue and provided possible solutions. The state of the art is a system called PrivSQL [15] which makes it possible to more efficiently produce tables via SQL commands while attempting to enforce a given privacy budget and while also attempting to optimize query accuracy. An architecture diagram for this system is shown in Figure 5-5. The system must generate a set of differentially private views for a set of preset queries. A sensitivity analysis must be performed and a set of protected synopses are then generated that can be publicly viewed. Census will perform the appropriate queries and create the protected tables using this

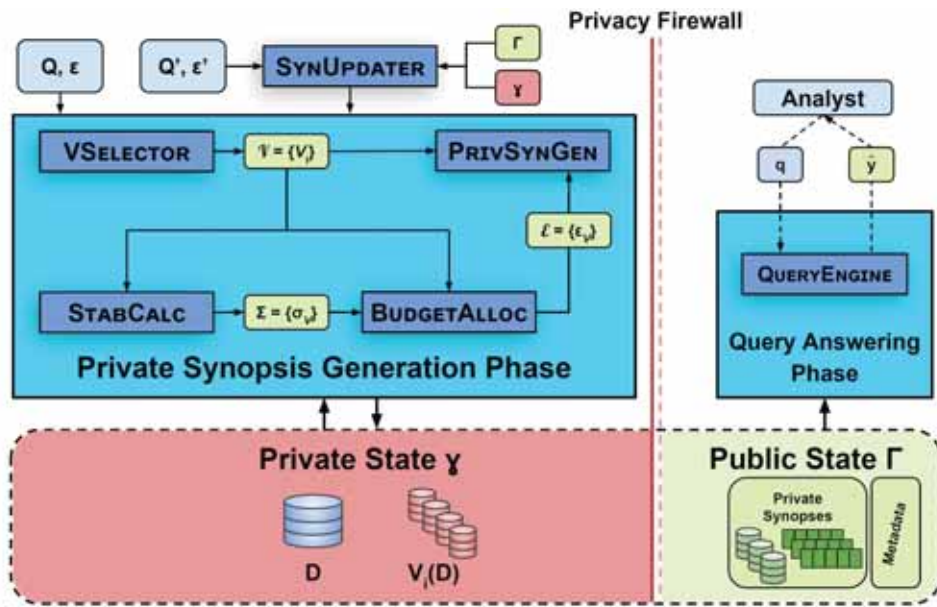


Figure 5-5: Architecture diagram for private SQL queries [15].

approach. Microdata associated with these tables will then be produced. This is at present work in progress, At the time Census briefed JASON their plan was to release a modified version of SF1 but tables requiring the linkage of data from person and housing records could not yet be constructed. It is expected that with further work using PrivSQL it should be possible to eventually produce many if not all of the traditional Census products.

## 5.5 The Dinur-Nissim Database under Differential Privacy

We provide here an example of the use of methods like DP as applied to queries of the Dinur-Nissim dataset. As discussed in Section 4.2 Dinur and Nissim made use of a simple database consisting of binary numbers to put forth what is now known as the Fundamental Law of Information Recovery, namely, that even in the presence of noise one can determine the contents of a private database by issuing and receiving the responses to too many queries. Here we illustrate that, despite the addition of noise, it is still possible to obtain meaningful statistical information from the database. We create a DN database as an array of randomly chosen

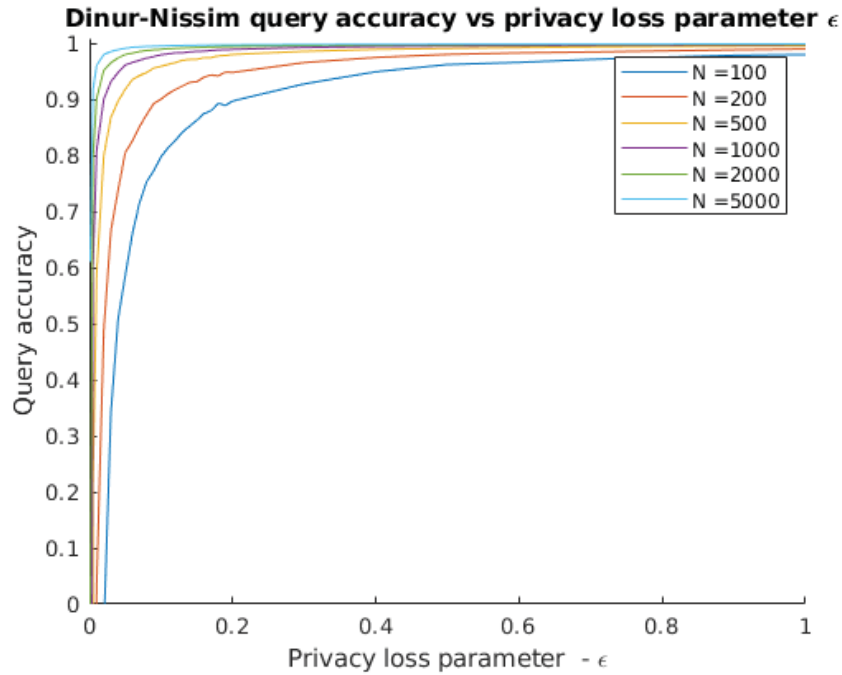


Figure 5-6: Accuracy of a sum query on the DN database. The values of  $N$  shown indicate the size of the database.

bits of size  $N$  bits. These could be the answer to a survey where the response is yes or no. We would like for example to know how many people responded yes to our survey. The result of our query is just the sum of the bits giving us the number of affirmative answers. For any query of this type issued we add a random amount of noise sampled from a Laplace distribution  $\text{Lap}(1/\epsilon)$  with mean zero and variance  $2/\epsilon^2$ . To measure the impact of the additional noise we calculate the query accuracy defined by

$$A = 1 - \frac{|\tilde{S} - S|}{S}$$

where  $\tilde{S}$  is the noised sum and  $S$  is the sum in the absence of noise.  $A$  varies from 1 (no error) and then decreases towards zero and can become negative. Clearly,  $A$  of zero is of no utility. For each value of  $\epsilon$  and  $N$  the number of bits we repeated the calculation 1000 times and reported the average  $A$ . The results are shown in Figure 5-6.

As can be seen, the accuracy of a query perturbed using the Laplace mech-

---

anism depends on the size of the data set. For the smallest dataset of size 100, a privacy loss value of  $\epsilon = 2$  degrades the query accuracy by about 15%. As  $N$  is increased the query accuracy improves and for  $N = 5000$  the effect of the perturbation due to DP is imperceptible. In fact it would be smaller in this case than the statistical uncertainty associated with the query which varies as  $1/\sqrt{N}$ . For smaller values of  $\epsilon$  the impact of the perturbation becomes more noticeable with the conclusion that smaller values of  $\epsilon$  that provide increased privacy protection will not disturb statistical accuracy provided one deals with large datasets.

## 5.6 Multiple Query Vulnerability

As discussed in section 4 for the Dinur-Nissim dataset, it is still possible to recover the bits of the dataset provided enough queries are issued and optimization is used to get a “best fit” to the bit values. This works in our case even in the presence of arbitrarily large noise. The optimization technique, in our case least squares with constraints followed by rounding, can apparently return a result that converges to the true answer - the values of the bits in the dataset. We note that the residual norm of the optimization in this case will be very large, indicating that when the optimized result is used to compute the right hand side of the linear system representing the queries, the difference with the right hand side presented to the optimizer is very large. This is to be expected as we constrain the lower and upper bounds of the solution to be zero and one respectively. When we apply, for example, Laplace noise to the right hand side, we perturb it so that in some cases it would be impossible for a series of zeros and ones to sum to the indicated right hand side values. The larger is the noise amplitude, the more likely this is to occur. Nevertheless the optimizer will find solutions (effectively averaging out the applied noise) and as the number of random queries is increased the percentage of recovered bits increases.

To put this observation into the context of the Census vulnerability, we generate a Dinur-Nissim database consisting of 4000 randomly chosen bits. We then

---

generate a query matrix  $Q$  of size  $N_Q \times n$  where  $n$  is the size of the database and  $N_Q$  is the number of issued random queries. In this case we set  $N_Q$  to be a multiple of the dataset size as this seemed more relevant to the issue faced by Census. That is, given a population, how many queries expressed as a multiple of the population suffice to infer the microdata. In the case of the Dinur-Nissim dataset, it is possible to ask this question even in the presence of noise and, empirically, while the number of queries required to determine the bits does increase with the size of the dataset, eventually, with high probability, all the bits can be recovered.

Given a query matrix and the dataset we compute the matrix-vector product and then set a value of the privacy loss parameter  $\epsilon$  (in our case ranging from 0.01 to 1) and added to each component of the vector a random amount of noise sampled from the Laplace distribution. We then applied constrained least squares optimization and examined the fraction of bits recovered correctly. We assume that different bit locations are recovered correctly in computing the fraction recovered, but privacy concerns would certainly arise if the fraction of bits recovered exceeded 0.9. After some number of queries the algorithm succeeds in determining all the bits every time. A Matlab code performing this computation is included in Appendix B.

The results of our experiment are shown in Figure 5-7. Note that if one just guesses randomly, it is possible to recover 50% of the bits and so the minimum fraction of bits recovered is 0.5. The  $x$ -axis of the plot (labeled "Query multiple") indicates the number of queries scaled as a multiple of the size of the data set. In this case a multiple of 20 indicates 80000 random queries were made. The  $y$  axis indicates the privacy loss parameter. It can be seen that for example for  $\epsilon = 0.01$  and 4000 queries the results are not much better than random. But as the number of queries increases the fraction of bits recovered also increases. As the privacy parameter increases, and the number of query multiples increases eventually all the bits are recovered. This behavior is in line with the results of DP. Not only must one noise the data, one must also restrict the number of queries.



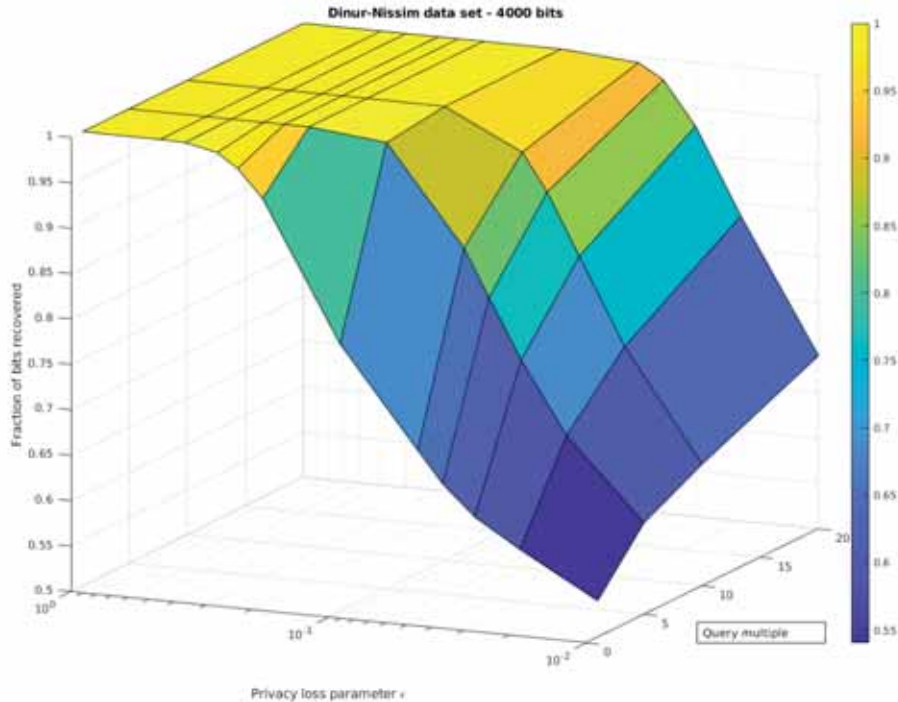


Figure 5-7: Fraction of bits recovered for the Dinur-Nissim database as a function of the privacy loss parameter and the number of multiples of the size of the database.

## 5.7 Disclosure Avoidance using Differential Privacy

The Census proposes to use an idea similar to that discussed above using the Dinur-Nissim database but applied to the much more complex microdata collected by the Census. As noted above, if one post-processes data that have been previously processed through an algorithm that satisfies the DP conditions, then the post-processed data will also satisfy the constraints of DP provided the original data are not accessed again during the post-processing.

If one creates the usual histograms as published by the Census (i.e. PL94, SF1, etc.) and then applies a DP mechanism to the results, then one could apply the same optimization technique used to demonstrate the Census vulnerability in Section 4 to produce microdata that are now themselves protected by DP. This

---

approach will create synthetic microdata upon which statistical queries can then be issued. We detail below the proposed approach following closely the briefing to JASON by Dan Kifer [14].

The approach Census will use has three phases

1. Select
2. Measure
3. Reconstruct

The microdata are first represented as a multidimensional histogram  $H$ . These are the tables that Census typically publishes. This histogram is then flattened into a column vector. A query on this histogram  $H$  is a linear function of the vector and can be represented by a query workload matrix  $Q$ . To acquire the answer to a prescribed set of queries we simply compute  $QH$ .

**Selection phase** In the selection phase a strategy matrix  $A$  is constructed for the purpose of optimizing the accuracy of various queries. A well chosen strategy matrix will minimize the sensitivity associated with the chosen queries by reducing the statistical variance of the queries. Algorithms for computing such a matrix are given in [20], but require some understanding of what the preferred query workload would be so that the appropriate set of queries is optimized for accuracy.

**Measurement phase** In this phase the query workload is performed with noise then added to the result. The amount of noise will depend on the sensitivity of the query and the chosen value of  $\epsilon$ :

$$\tilde{Y} = AH + \text{Lap}\{\Delta_A/\epsilon\}$$

where  $\tilde{Y}$  is the DP response to the query and  $\Delta_A$  is a norm measuring the sensitivity of the strategy matrix  $A$ .

---

**Reconstruct** The final step is to estimate  $QH$  from the vector  $\tilde{Y}$ . This requires undoing the multiplication by the strategy matrix:

$$QH = QA^+\tilde{Y}$$

As the strategy matrix may not be square, the Moore-Penrose pseudo-inverse is used to compute  $H$  and then  $QH$ .

The measurement phase consumes the privacy budget. Once this is accomplished the results could in principle be released to the public. The reconstruction phase will not re-access the private data and hence does not require additional privacy budget. The cleverness of this idea is that the final product can even be in the form of microdata which can then be reprocessed by users of the Census data. What is less clear however, is the accuracy of queries that have not been optimized using the High Dimensional Matrix Method and whether the results of those queries will have an acceptable statistical utility. This will be discussed further in Section 6.

While the steps of this procedure are easily described, the computational aspects of doing this for the census pose significant challenges. Recall that for the country Census publishes billions of queries and so the histogram will have billions of cells. The query matrix could be as large as the square of the histogram size depending on what measurements are to be reported. Choosing a strategy matrix based on the potential query workload is not feasible. The reconstruction is also going to entail an enormous computational cost as a result of the matrix sizes. Finally, the result of the multiplication by the Moore-Penrose inverse will lead to non-integer results. If we wish to convert these to sensible microdata a second phase will be required in which the results of the first phase will have to be converted to integers. Once this is done the optimization approach taken by Census to reconstruct the microdata can be used to create differentially private microdata.

The solution to the challenges discussed above are to break the problem up into pieces and then perform the DP reconstruction on each piece. The first

---

attempt to do this was a “Bottom Up” approach in which the select-measure-reconstruct approach was applied to each Census block and then converted to microdata. This has the advantage that the operations are all independent for each block and the privacy budget is simple - one value of  $\epsilon$  can be assigned to each block. The privacy cost does not depend on the number of blocks as each of these is processed independently of the others. It also has the advantage that the counts at various levels of the Census hierarchy are consistent. However, the injection of the DP noise adds up as the data are combined to form results for block groups, tracts, etc. A county in a populous region that contains many blocks will have an error proportional to the number of blocks. The “Bottom Up” approach is easy to conceptualize but it doesn’t use the privacy budget efficiently.

Instead, Census will use a “Top-Down” approach. The privacy budget is split into six parts: national, state, county, tract, block group and block. A national histogram  $\tilde{H}^0$  is then created using the select measure and reconstruct algorithm outlined above. This involves the population of the US but the number of queries is now manageable as the queries are not specified over geographic levels finer than the nation. Once this protected histogram is in place the same process can then be applied for the states using the privacy budget allocated for states. These histograms are constrained so that they are consistent with national totals. This process is then followed down to the county, block group and finally the block level. Once a protected histogram with non-negative integer entries is created it can then be transformed to microdata using the optimization approach Census used to determine the reconstruction vulnerability as discussed in Section 4. The Top-Down approach has the advantage that it can be performed in parallel and the selection of queries can be optimized at each level making it possible to use the privacy budget more efficiently. It also has the advantage that it enforces any sparsity associated with 0 populations at various levels (for example someone over 100 who indicates they are a member of five racial categories). These are known as structural zeros.

In producing an appropriate histogram that can be turned into microdata two

---

optimizations are performed. The first is a least squares optimization which effects the Moore-Penrose inverse subject to various constraints that the histogram being determined must be consistent with the parent histogram. For example the total population of the states must sum to the population of the country. The result of this optimization leads to fractional entries and so the second step is to perform an optimization that assigns integer values to the histogram cells such that the entries are non-negative integers that are rounded values of the fractional results and that sum to the same totals consistent with the parent histograms. This “rounding” step is performed using the Gurobi solver [12].

A complication in executing the TopDown algorithm is the need to publish some data without protection. These correspond to the invariants discussed in Section 5.3. Census plans to provide accurate counts of the population of each of the 50 states, DC and Puerto Rico to support apportionment of Congressional representatives. It might also be desirable to report correct population down to the census block.

But in addition, there are other constraints and so it would be desirable to be consistent with these. For example, the number of occupied group quarters and housing units in each census block is public information as a result of a program called Local Update of Census Addresses (LUCA). This program is used by Census to update the Master Address File (MAF) used to distribute census surveys. The addresses themselves are protected under Title 13 but the number of group quarters is publicly released. As a result, if a census block were to have an occupied jail then the TopDown algorithm must assign at least one person to that jail. As another example, the number of householders in a block should be at least the number of households [14]. There are other data-independent constraints. For example, if a household has only one person in it then that person is presumably the householder.

Census has proposed a partial solution to this problem by casting the constraints as a series of network flows that can then be appended to both the least

---

squares and rounding optimizations described above [14]. This work is still experimental at the time of this writing and will be further evaluated.

The enforcement of invariants such as national and state populations presents no issues in terms of the DP computation. Neither does the enforcement of structural zeroes such as there cannot be any males in a dormitory that is all female. But the constraints that are independent of the data such as the fact that a grandparent must be older than the children in a household creates issues of infeasibility as the optimization recurses down the Census geographic hierarchy. If such implied constraints are ignored there is the possibility that for example assignments at the block group level are not consistent when extended to the higher Census tract level. When this happens it is called a “failed solve” and Census then applies a “failsafe” optimization. The constraints impeding solution are relaxed and the optimizer finds the closest feasible solution meaning a violation of the exact constraint will be allowed. The assignments at the higher geographic level (for example the county level of optimization at the tract level fails) are then modified to maintain hierarchical consistency. The overall impact of the use of the failsafe on the utility of the protected Census data is still not fully understood and is an area of ongoing research. One approach that would avoid this difficulty is to not insist on hierarchical consistency at the finer geographic levels, in particular census blocks. For example providing the correct population in each block might not be enforced as a constraint. This however may have implications for the use of census data in the redistricting process, an issue we discuss in Section 6.

The new disclosure avoidance scheme will now look as in Figure 5-8. It is expected that Census will still perform the usual imputations associated with households and general quarters for which Census enumerators cannot obtain information but, at present, no household swapping will be performed. Instead the Census will apply the TopDown algorithm and then create a set of noised tabular summaries and also, for the first time, the synthetic microdata associated with the summaries.

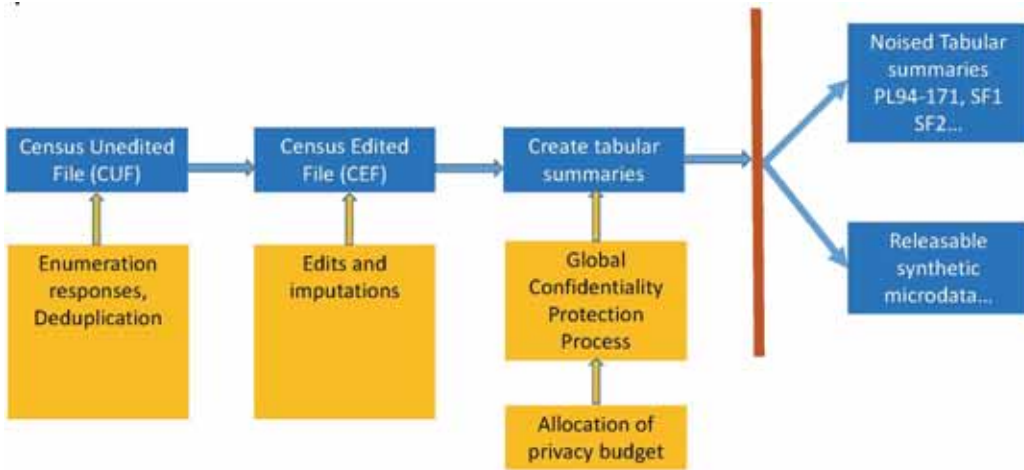


Figure 5-8: A graphical representation of the proposed DAS using the TopDown DP algorithm.

The proposed disclosure avoidance system using DP has been implemented in Python and is publicly available [34]. Work continues to improve query accuracy and enforce invariants and implied constraints. Census is to be commended for making this software available to the community so that it can be examined in detail and inform users on the details of the application of DP to census data.

---

## 6 ASSESSING THE ACCURACY-PRIVACY TRADE-OFF

In this section we examine the results of some of the early applications of the new Census DAS on census data. As mentioned in Section 5 Census has publicly released the DAS software. To further aid users, it has processed census data from 1940 and produced synthetic microdata. It has also released some preliminary assessments of query accuracy for the 2010 census data. We discuss these results here with an emphasis on the trade-off between query accuracy and the level of privacy protection.

### 6.1 Census Analysis of 2010 Census Data

Census has applied the proposed DAS using DP to the 2010 census data. The advantage here is that the schema for the 2010 census largely overlap with the schema for the forthcoming 2020 census. But a disadvantage is that this data is not yet publicly available. By law census data can only be publicly released no earlier than 72 years after a census is taken so the latest data available to the public is the 1940 census. We are able to provide only a limited view of the results of the Census analyses on 2010 data as most of these are not yet available for release and are still protected under Title 13. JASON did have access to these results but the assessment provided here can only describe them qualitatively.

As briefed to JASON by P. LeClerc [16], Census has executed the TopDown algorithm on a histogram from the Census Edited File  $H_{CEF}$  to produce a noised histogram of privatized results  $H_{DAS}$ . The experiments were performed for the PL94-CVAP product that has 4032 entries representing a shape of  $8 \times 2 \times 2 \times 63 \times 2$ . Recall that this product is used to examine voting districts to ensure adherence to the Voting Rights Act and includes the following pieces of information:

- 8 group quarters-housing units levels,



- 
- 2 voting age levels,
  - 2 Hispanic levels,
  - 63 OMB race combinations,
  - 2 Citizenship levels.

For each state one can create such a histogram and examine it at various geographic levels: state, county, tract, block group and block. For each geographic level (geolevel)  $\gamma$ , Census executed 25 trials of the DAS, averaged over the results, and reported a number of metrics. We will consider here only one of them:

$$\text{TVD}_\gamma = 1 - \frac{L^1(H_{DAS,\gamma}, H_{CEF,\gamma})}{2\text{POP}_\gamma}.$$

This can be thought of as a type of accuracy metric using the  $L_1$  norm or sum of the magnitudes of the distance between the DAS and CES entries. This is similar in some respects to the Dinur-Nissim query accuracy metric discussed in Section 5.5. If the DAS and CEF histograms were to agree across all components at a given geographic hierarchy level  $\gamma$ , the TVD value would be exactly 1. The possible difference between the values is normalized by twice the population, but this does not provide an absolute lower bound on the TVD metric and it can become negative depending on how much noise is infused into the histogram values.

As of the date of this report, Census has publicly released TVD metrics for the state of New Mexico [30]. These indicate query accuracy vs. privacy loss for actual Census data and may be reflective of the results of the future 2020 Census. In Figure 6-1, the TVD metric as a function of  $\epsilon$  is plotted at the state, county, tract group, tract, block group and block for the state population. As  $\epsilon$  increases from 0, the TVD metric will tend to one indicating that as  $\epsilon$  increases less noise is injected into the histograms until at sufficiently large  $\epsilon$  the DAS and CEF results agree in this norm. As can be seen, for geolevels with large populations (e.g. counties, tracts and even block groups) the TVD metric for population is close to one for values of  $\epsilon$  as small as 1/2. At even lower levels of  $\epsilon$  we see the same

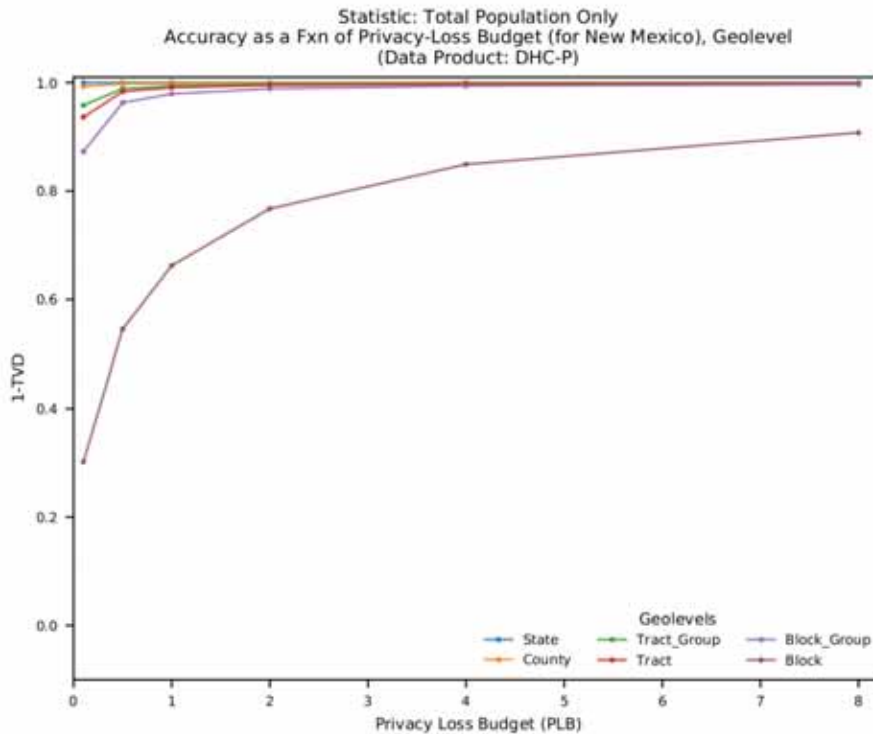


Figure 6-1: A plot of the TVD metric for total population for various geolevels as a function of privacy loss parameter for the state of New Mexico [30].

type of degradation of query accuracy as in the Dinur-Nissim example. Because we cannot tie TVD to a measure of statistical accuracy we cannot comment on whether such degradation of accuracy would or would not be acceptable from that point of view. At the block level, because populations are typically much smaller than block groups the degradation is noticeable and even at  $\epsilon = 4$  we still have  $TVD \approx 0.8$ .

In Figure 6-2 we show again the TVD metric but this time for a subhistogram looking only at those entries associated with race and Hispanic origin. Typically the counts here will be smaller particularly as we examine the finest block level and so the TVD metric deviates further from 1 than shown in Figure 6-1 as the privacy loss budget is decreased.

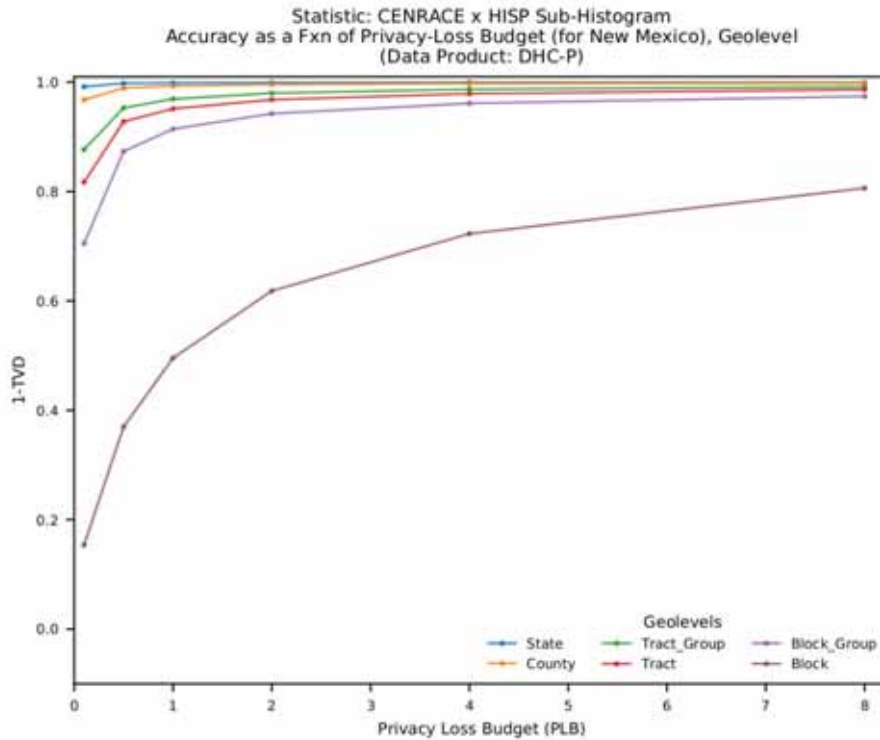


Figure 6-2: A plot of the TVD metric for race and Hispanic origin for various geolevels as a function of privacy loss parameter for the state of New Mexico [30].

The TVD metric provides some insight into the degradation of query accuracy as the privacy loss budget is decreased, but it suffers from being a coarse measure of accuracy as it sums over the entries at a given geolevel and so does not provide a view of the variance of the individual differences. For example, it would be useful to see the distribution of TVD measure block by block. A more detailed assessment in terms of microdata but for the older 1940 Census is discussed in the next section.

## 6.2 IPUMS Analysis of 1940 Census Data under the Census DAS

IPUMS (Integrated Public Use Microdata Series) is an organization under the University of Minnesota Population Center providing census and survey data from a

---

variety of countries. It is the world's largest repository of census microdata. JASON was briefed by Dave van Riper of IPUMS [36] (cf. also [37]) who examined in detail the application of the Census DAS to the 1940 Census microdata. We note that JASON has not verified this work but we discuss it here to give examples of the differences between counts associated with the DAS processed synthetic microdata and the true census microdata. As discussed in Section 4, we expect more dispersion as we descend to finer geographic regions. At the time of van Riper's briefing he had performed comparisons for Minnesota census data. Since then, he has also performed analyses for the entire US and it is this data that we discuss here.

It should be noted that the geographical hierarchy for the 1940 census was different than that used today. The finest level of geographic resolution is what was then called an enumeration district. Enumeration districts are roughly comparable to census block groups on the geographic spine and also similar in some ways to what Census terms "places". The median population for enumeration districts was about 1000 people. The median population for census places in 1940 was about 800 people.

As indicated in Section 5, Census has publicly released differentially private microdata for the 1940 census. Microdata files were generated for the entire country for eight different values of the privacy loss parameter  $\epsilon$  : 0.25, 0.5, 0.75, 1.0, 2.0, 4.0, 6.0, 8.0. Four runs of the DAS were provided at each value of  $\epsilon$ . The microdata made available are those of the PL94-CVAP Census product and include whether a respondent is of voting age, Hispanic origin and Race as well as household and group quarters type at four geographic levels: national, state, county and enumeration district. IPUMS did not run the Census DAS to generate synthetic microdata. Instead it analyzed those results generated by Census to compare against unfiltered microdata that constitute ground truth. The source code for the DAS system [34] is configurable so that one can allocate fractions of the total privacy budget over the various geographic levels and tables. In this case the budget is allocated evenly over geographic levels. Each level of the hierar-

---

chy receives a quarter of the total privacy budget. Allocations must also be made for the various tables that are produced and then subsequently noised by the DP algorithm. In this case Census chose the following fractions:

- Voting age by Hispanic Origin by Race: 0.675
- Household group quarters type: 0.225
- Full cross of all variables: 0.1

The fraction of the total privacy budget to be allocated for each level and for each table is then the product of the geolevel allocation times the table fractions. For a given total privacy loss budget  $\epsilon$  it is these fractions that are used to provide the noise levels for each individual table at a given geographic level. For example if the total privacy budget were 0.25 then the privacy budget for each histogram will look as shown in Table 6-3. The table shows the effective values of  $\epsilon$  but also the level of dispersion for an equivalent Laplace distribution. These dispersion levels will affect various tables differently. A table associated with large counts will not be significantly affected by an  $\epsilon$  corresponding to a dispersion of 300 but a table at the enumeration district level could be significantly affected.

Box plots of the distribution of populations across all US counties in 1940 are shown in Figure 6-3 for all the values of  $\epsilon$  used in the Census runs of the DAS. The distribution as computed by IPUMS from the true 1940 microdata is shown at the left of the Figure. As can be seen, as  $\epsilon$  increases the box plots converge to the IPUMS result. For the lowest value of  $\epsilon$  used, differences can be seen for populations of 100 or more. By and large, the box plots are quite similar across the various values of  $\epsilon$ . More insight into the effect of the DAS at the finer geolevels can be seen in Figure 6-4 where box plots for the differences between the DAS and IPUMS population estimates are shown. The orange box plots represent counties and the teal plots represent enumeration districts. Again as  $\epsilon$  increases we see the differences reduce. But at lower values of  $\epsilon$  differences on the order of several hundred people appear when we look at various outliers. It should be noted

Geography level	Table	$\epsilon$ for Table at level	Noise dispersion
Nation	Vot-Hisp-Race	0.042	47.4
Nation	HouseholdGenQuart	0.014	142
Nation	Detailed	0.006	320
State	Vot-Hisp-Race	0.042	47.4
State	HouseholdGenQuart	0.014	142
State	Detailed	0.006	320
County	Vot-Hisp-Race	0.042	47.4
County	HouseholdGenQuart	0.014	142
County	Detailed	0.006	320
Enum Dist	Vot-Hisp-Race	0.042	47.4
Enum Dist	HouseholdGenQuart	0.014	142
Enum Dist	Detailed	0.006	320

Table 6-3: Values of the privacy budget allocated to the various geolevels and tables by the Census DAS system for the 1940 Census data [36]. The noise dispersion is listed here to give some notion of the variance of the noise applied to the data. In this case the value  $\epsilon = 0.25$  is used [36]

that the box plots are not normalized and that the teal box plots for enumeration districts are smaller simply by virtue of representing smaller populations.

Van Riper has also computed how the populations of counties compare in detail in Figure 6-5. The Figure plots the IPUMS value for a county population vs. the DAS value. The level of agreement is measured by how closely the two values would lie to the 45° line indicating equality. As can be seen the county populations align well at all values of  $\epsilon$ . In contrast, for enumeration districts we see in Figure 6-6 more dispersion. This is most observable as  $\epsilon$  becomes smaller. Note that because the DAS does not allow negative population there is a pile-up as population size decreases. Such results are to be expected as one focuses on finer geolevels and smaller populations.

The same analysis has been performed for population under 18 across all US counties for the 1940 Census. These are shown in Figure 6-7. This too looks quite

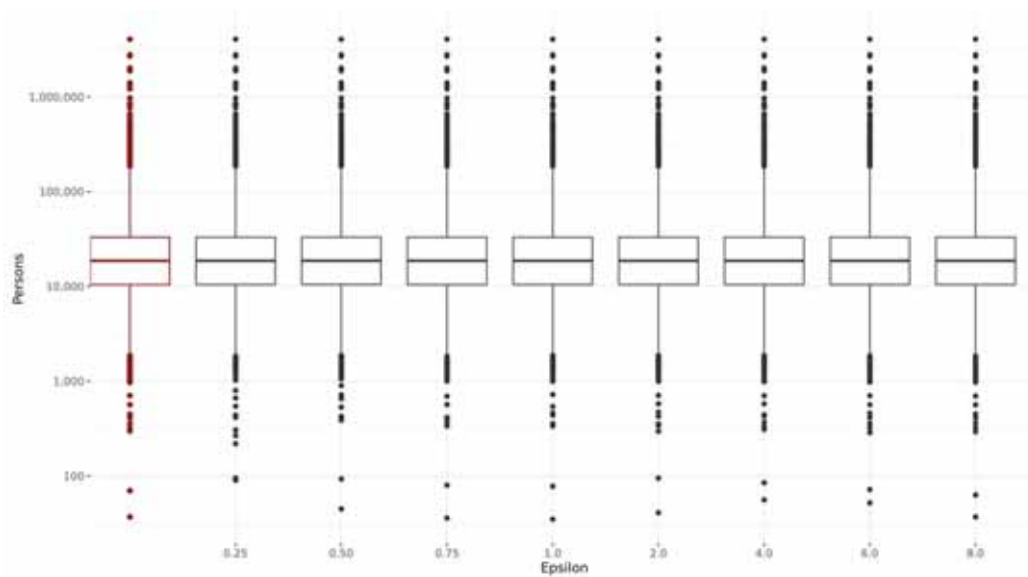


Figure 6-3: Box plots for the distribution of total US population in 1940 under different values of the privacy loss parameter [36].

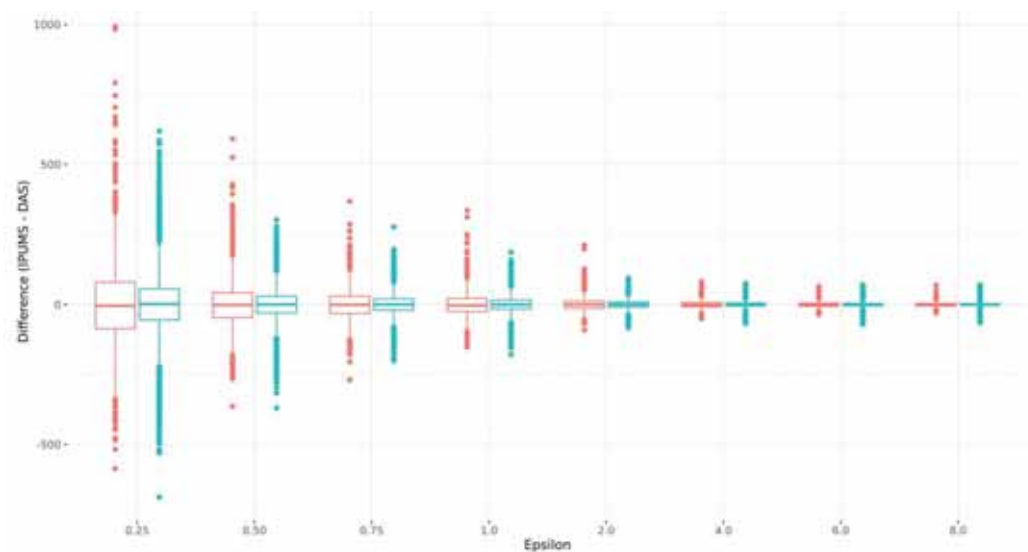


Figure 6-4: Box plots for the differences between IPUMS and Census DAS for total population counts under different values of the privacy loss parameter [36].

similar to population estimates with some issues seen for counties with smaller populations at lower values of  $\epsilon$ . The corresponding results for enumeration districts are shown in Figure 6-8. Because we are now focusing on a subgroup of the population for enumeration districts there is yet more dispersion in the results. But

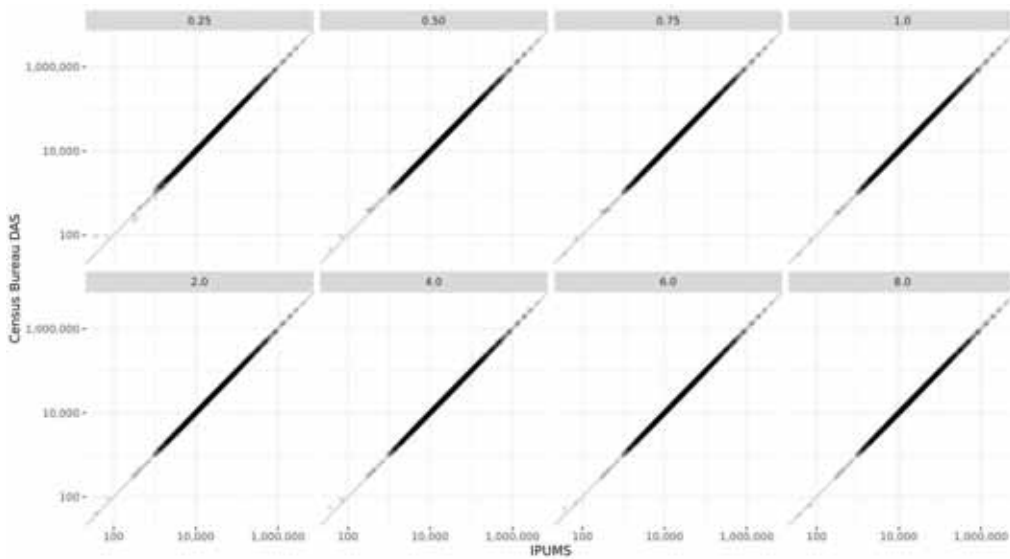


Figure 6-5: Total population for US counties under differing levels of the privacy loss parameter [36].

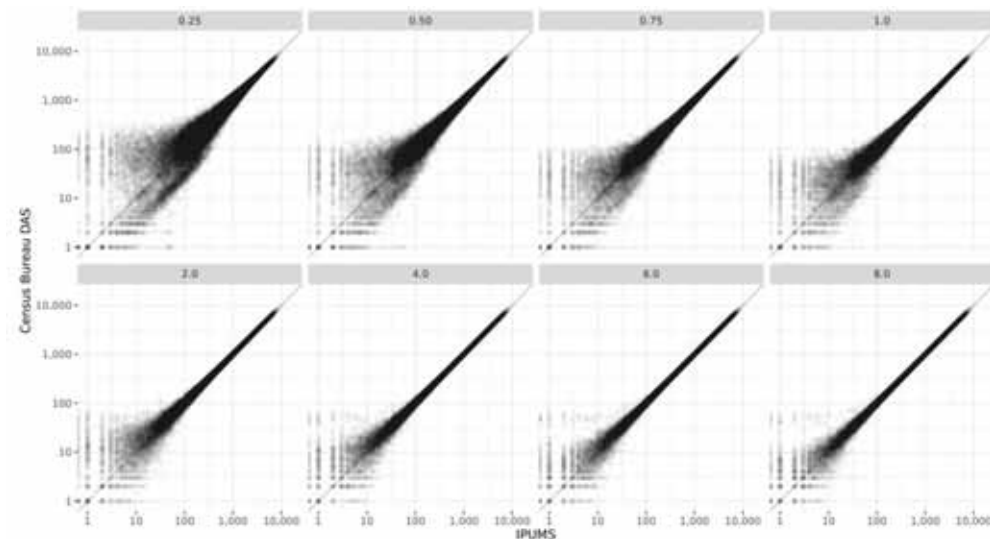


Figure 6-6: Total population for US enumeration districts under differing levels of the privacy loss parameter [36].

perhaps of some concern is that in some enumeration districts the DAS indicates a large number of people under 18 when there are in fact very few. There are some enumeration districts with 50 or more people where this particular application of the DAS (with values of  $\epsilon$  of 0.25, 0.5 and even in some cases 1.0) indicates that 100% of the population is under 18, an observation that could have implications



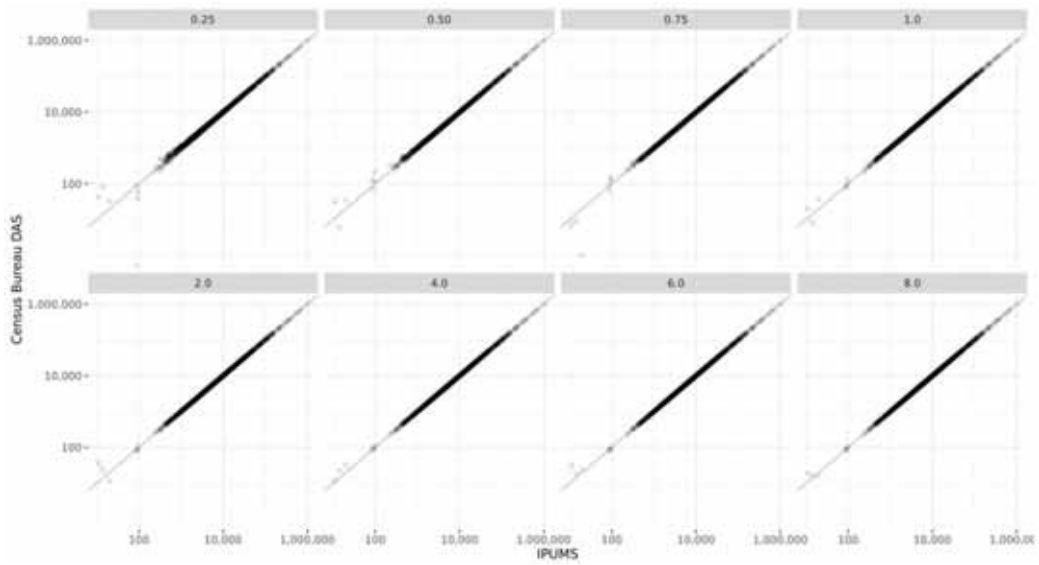


Figure 6-7: Total population under 18 for US counties under differing levels of the privacy loss parameter [36]

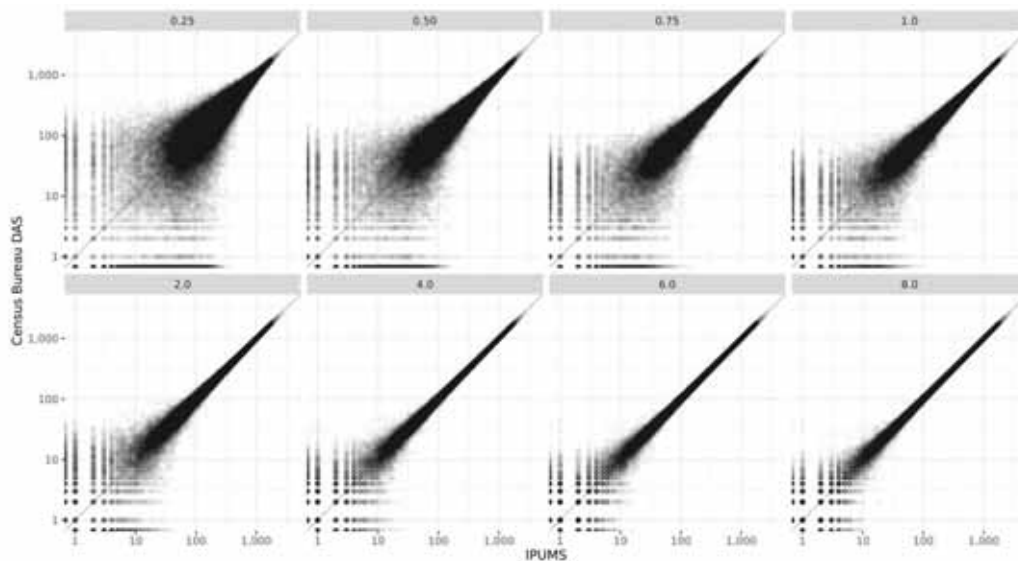


Figure 6-8: Total population under 18 for enumeration districts under differing levels of the privacy loss parameter [36]

for assessments of voting age population, a component of the information needed for the PL94 publication.

Several points should be emphasized in examining the current application of the DAS:

- 
- The DAS does not unduly perturb statistics at the national, state and even largely at the county level at all the values of  $\epsilon$  considered.
  - The dispersion seen in the IPUMS-DAS comparison for enumeration districts is to be expected at lower values of  $\epsilon$ . The DAS is after all meant to protect small populations.
  - The application of the DAS will degrade the utility of various statistics. This degradation will increase as one further restricts the population by characteristics such as race, voting age, etc. This illustrates a trade-off inherent in the use of DP among privacy, accuracy and granularity of queries. The requirements for accuracy will need to be determined in the future through consultation with external users of the data. We discuss this trade-off further in Section 7.
  - The allocation of the privacy budget can be modified depending on the accuracy requirements. For example it would be possible to allow for larger privacy loss parameters for some tables and less for others provided the total privacy budget is conserved.
  - The current version of the DAS is a demonstration product. For example, at the time of this writing, the implementation presented here does not benefit from the improved accuracy of the high dimensional matrix method. Nor do the products contain all the invariants and constraints that the Census bureau has identified. Work is in progress to improve query accuracy to the extent possible. As these improvements are made it will be important to continue to reevaluate the performance of the DAS against ground truth.



---

## 7 MANAGING THE TRADE-OFF OF ACCURACY, GRANULARITY AND PRIVACY

Published census tabulations must balance inconsistent desiderata. They should be accurate (i.e., published counts should be the sums of the underlying micro-data). But tabulations should also be appropriately granular (i.e., have a high level of detail such as block, gender, age, race/ethnicity, etc. But, as has been discussed, pushing granularity to the extreme can create small (or even singleton) counts in table entries (particularly in small blocks), thereby eroding privacy. Of course, privacy could be enhanced and granularity preserved by relaxing the accuracy requirement (as embodied in DP or swapping schemes). Alternatively, privacy could be enhanced and accuracy preserved by reducing granularity. The situation can be illustrated by the “disclosure triangle”, where the balance among the three competing considerations of privacy, accuracy, and granularity varies across the interior as shown in Figure 7-1.

No compromise will be perfect. In this section, we discuss some aspects of managing this trade-off.

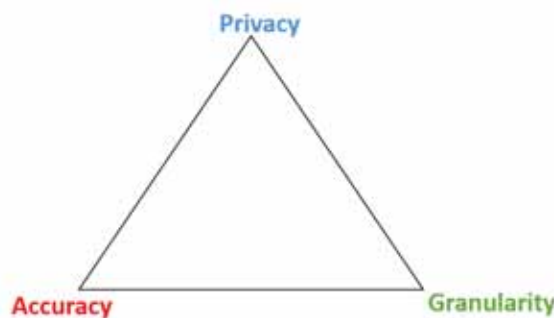


Figure 7-1: Census must balance, accuracy, granularity and privacy in its publications. It is not possible to achieve all three simultaneously.

---

## 7.1 Risk Assessment

The use of DP is clearly promising as a way to protect census data, but it is important to recall the original motivation for its use. Its proposed use was primarily motivated by the 17% re-identification rate assessed by Census using the 2010 tables, and thus the degree to which DP prevents re-identification needs to be similarly explored. Technically, differential privacy as pointed out by Reiter [28] is a guarantee

“on the incremental disclosure risks of participating (in a survey) over whatever disclosure risks the data subjects face even if they do not participate (in the survey)”.

It does not provide an assessment of disclosure risk in and of itself. It is also not one methodology. A number of algorithms can be applied and must be implemented correctly. In the case of its use for the census there are clearly complications like invariants, implied constraints etc. that will require further work and assessment. For these reasons, explicit quantification of the risk of re-identification is still required. The choice of  $\epsilon$  should be informed by calculations of the risk of re-identification using the methods developed by Census and linking with current commercially-available data but applied to microdata as processed through DP. JASON understands that this will be significantly more difficult than the original analysis that led to the re-identification of the 2010 Census data vulnerability. This is because the matching of the microdata in the absence of noise to commercial data was aided by the availability of the geographic location. The synthetic data generated by DP algorithms will not have this feature and so matching to commercial data bases will have to be performed using probabilistic record linkage (cf. for example [9]). A very useful property of DP here is that such linkage can be attempted at various values of  $\epsilon$ . At very high values of  $\epsilon$  we expect to recover the noise-free values and so we would also verify the previously assessed re-identification level of 17% against commercial marketing databases. But as  $\epsilon$

---

is decreased this re-identification rate must degrade. An open question is at what value of  $\epsilon$  would it degrade to a value sufficiently low so as to be administratively acceptable? While no official value of such a lower bound has ever been provided (nor would we expect one to be) presentations from Census have indicated that the re-identification rate of 17% was viewed as something like four orders of magnitude higher than previously assessed [27].

The fact that methods of data science will improve and commercially available data will become more comprehensive over time does not obviate the need for an analysis that can inform the current decision. Knowing the outcomes based on current data can help to support a choice of  $\epsilon$ . Once some assessment of an appropriate “upper bound” for  $\epsilon$  based on disclosure risk is in hand, further considerations regarding statistical accuracy for future queries on the data can be made in ultimately deciding the level of noise to be applied to the 2020 data.

## 7.2 Engaging the User Community

Analyses of aggregate data involving large populations will be minimally impacted by DP. Impacts will increase as one focuses on finer levels of geography or other demographic measures. We emphasize that this is precisely the desired impact of DP because individuals within a smaller group will be more identifiable, and thus it is precisely this “blurring” from DP that protects the privacy of these individuals. This aspect of DP needs to be effectively communicated to future users of Census data.

The challenge is to better quantify the balance of privacy protection and data utility for smaller groups. There are multiple communities with a deep interest in the accuracy-privacy-granularity tradeoff:

**State governments and redistricting commissions** These bodies are responsible for the drawing of Congressional and State legislative districts. PL94-171 requires the Census to provide to these bodies an opportunity to identify

---

the geographic areas relevant to redistricting and to then deliver tabulations of the population as well as race, race for population 18 and over (voting age), Hispanicity and Hispanicity for those 18 and over, occupancy status and, in 2020, group quarters population by group quarters type.

**Local governments** Local governments use census data for redistricting as well as to inform assessments of public health, safety, and emergency preparedness for the residents.

**Residents** Residents use census data to support community initiatives and to decide where to live, learn, work and play.

**Social scientists and economists** Census data forms a foundation for demographic studies as well as economic research.

Census has to some extent reached out to these communities through a July 2018 Federal Register Notice as well as several academic conferences [23]. The feedback received by Census emphasized several aspects:

- There was little understanding as to the need for application of Differential Privacy
- Users were vocal about the need to maintain block level data so that custom geographies could be constructed.
- Concerns were voiced about the potential loss of information for small geographic areas.

Clearly more work is needed and Census should participate actively in various fora, working with the community to characterize the scales and types of queries that will and will not be substantially impacted at different values of  $\epsilon$ . For example, opportunities for stakeholders to assess accuracy of queries on 2010 census data made available at various levels of protection would go a long way towards helping users assess the impact of DP on future analyses. In general it

---

will be necessary to engage and educate the various communities of stakeholders so that they can fully understand the implications (and the need for) DP. These engagements should be two-way conversations so that the Census Bureau can understand the breadth of requirements for census data, and stakeholders can in turn more fully appreciate the need for confidentiality protection in the present era of “big data”, and perhaps also be reassured that their statistical needs can still be met.

### **7.3 Possible Impacts on Redistricting**

As indicated above, redistricting bodies will require population and other data for regions with populations infused with noise from the DP process. There is concern that the population estimates derived from differentially protected Census block data will lead to uncertainties in designing state and Congressional voting districts. Census has begun to consider these issues, for example, in their recent end-to-end test for the state of Rhode Island [40]. We cannot discuss the variance of the actual counts and those treated under DP quantitatively here as these data are protected under Title 13. But, especially for the counts associated with smaller state legislature districts, the variances may lead to concerns in verifying that the districts are properly sized relative to the requirements of the Voting Rights Act. JASON was briefed by Justin Levitt [18] that such district equalization is a “legal fiction” since it is impossible to guarantee the accuracy and precision of the counts; they are a snapshot in time and so are not temporally static. Overall, the noise from block-level estimates is not expected to lead to legal jeopardy, but could in the case where, for example, racial makeup nears thresholds that elicit concern. Census is currently engaged with the Department of Justice regarding this issue but at the time of the writing of this report, Census has not allayed the Department of Justice’s concerns regarding this issue.



---

## 7.4 Limiting Release of Small Scale Data

The trade-off between probability of re-identification and statistical accuracy is reflected in the choice of the DP privacy-loss parameter. A low value increases the level of injected noise (and thus also decreases probability of re-identification) but degrades statistical calculations. Another factor that also influences the choice of privacy-loss parameter is the number and geographical resolution of the tables released, an aspect of granularity of the allowed queries. For example, if no block-level data were publicly released, a re-identification “attack” of the sort described above presumably would become more difficult, perhaps making it feasible to add less noise and so allowing a larger value of  $\epsilon$ .

For those public officials and researchers needing access to the finer scale block level data, special channels in the form of protected enclaves may be required. We discuss this next in Section 7.5. This most likely cannot be a solution for certain uses of Census data mandated by law. For example, redistricting must be performed in a way that is transparent to the public. Today this requires using block level populations in designing the new districts. These will be infused with noise under differential privacy. While it is thought that these population estimates can still be used for redistricting, their overall utility is closely tied to the value of  $\epsilon$  that is ultimately chosen. Too low a value of  $\epsilon$  may lead to concern over the totals. This seems to be a particularly difficult problem that must be solved in close consultation with the relevant stakeholders.

## 7.5 The Need for Special Channels

Depending on the ultimate level of privacy protection that is applied for the 2020 census, some stakeholders may need access to more accurate data. A benefit of DP is that products can be generated at various levels of protection depending on the level of statistical accuracy required. The privacy-loss parameter can be viewed as a type of knob by which higher settings lead to less protection but more

---

accuracy. However, products publicly released with too low a level of protection will again raise the risk of re-identification.

One approach might be to use technology (e.g. virtual machines, secure computation platforms etc.) to create protected data enclaves that allow access to trusted stakeholders of census data at lower levels of privacy protection. Inappropriate disclosure of such data could still be legally enjoined via the use of binding non-disclosure agreements such as those currently in Title 13. This idea is similar to the concept of “need to know” used in environments handling classified information. In some cases there may emerge a need to communicate to various trusted parties census data either with no infused noise or perhaps less infused noise than applied for the public release of the 2020 census. Examples include the need to obtain accurate statistics associated with state or local government initiatives, or to perform socio-economic research associated with small populations.

At present, the only way to obtain data not infused with noise is to apply for access via a Federal Statistical Research Data Center. These centers are partnerships between federal statistical agencies like the Census and various research institutions. The facilities provide secure access to microdata for the purposes of statistical research. As of January 2018, there were 294 approved active projects with Census accounting for over half of these. All researchers must at present obtain Census Special Sworn Status (to uphold Title 13), pass a background check and develop a proposal in collaboration with a Census researcher.

The use of DP presents an opportunity to expand the number of people who may access more finely-grained data but who would not need to access the original microdata. Products could be constructed at higher levels of the privacy loss parameter than that used in releasing Census data to the public. In a sense, the use of DP allows Census to control the level of detail available to a researcher but in accord with the users “need to know”, or more appropriately their need to access data at a given level of fidelity.

If such a program is developed there may arise the need to increase the ca-

---

capacity of the research data centers but at the same time the requisite security must be enforced. The defense and intelligence communities are facing similar issues and have responded by using cloud-based infrastructure and “thin client” terminals with limited input/output capability and strongly encrypted communication to ensure that data is appropriately protected and not handled improperly.

Transformative work in various areas of social science and economics has resulted from the ability to access and analyze detailed Census data. For example, Chetty and his colleagues [3] have used detailed census data to research approaches to using DP in small areas while maintaining the guarantees of DP. The development of virtual enclaves would expand opportunities to make similar contributions to a much wider cohort of researchers.

---

## 8 Conclusion

We conclude this report with a discussion of the controversy that has arisen as a result of the discovery of the Census vulnerability. The need to address the Census vulnerability also brings forward aspects of a tension between laws that protect privacy as opposed to those that require the government to report accurate statistics. We close with a set of findings and recommendations.

### 8.1 The Census Vulnerability Raises Real Privacy Issues

In the view of JASON, Census has convincingly demonstrated the existence of a vulnerability that census respondents can be re-identified through the process of reconstructing microdata from the decennial census tabular data and linking that data to databases containing similar information that can identify the respondent. The re-identification relied on matching Census records with commercial marketing datasets. These data providers, such as Experian, ConsumerView, and others already have a good deal of the data Census must secure such as name, age, gender, address, number in household, as well as credit histories, auto ownership, purchasing, consumer tastes, political attitudes, etc. But we note that the accuracy and granularity of their data is almost surely less than Census, and they generally do not include race or Hispanic identity; the latter is most likely a choice, not a fundamental constraint on information collection. In addition to this data there is also proprietary data maintained by Facebook, the location data collected by cell phone providers, etc.

One might argue that Census data is not of much additional utility given the limited amount of information gathered in the decennial census. However, many components of the data Census collects are not in the public domain and are still viewed as private information. For example information on children is hard to purchase commercially because its collection is enjoined by laws such as the Children's Online Privacy Protection Act. Other examples include race, number

---

and ages of children, sexuality of household members and, in the near future, citizenship status. Census has an obligation to protect this information under Title 13 and, in view of the demonstrated vulnerability, it is clear that the usual approaches to disclosure avoidance such as swapping, top and bottom coding, etc. are inadequate. The proposal to use Differential Privacy to protect personal data is promising although further work is required as this report points out.

The decision to use Differential Privacy has elicited concerns from demographers and social scientists. Ruggles has argued, for example, that Census has not demonstrated that the vulnerability it discovered is as serious as claimed. In [29] he states

“In the end only 50% of the reconstructed cases accurately matched a case from the HDF source data. In the great majority of the mismatched cases, the errors results from a discrepancy in age. Given the 50% error rate, it is not justifiable to describe the microdata as ‘accurately reconstructed’.”

Reconstructing microdata from tabular data does not by itself allow identification of respondents allow identification of respondents; to determine who the individuals actually are, one would then have to match their characteristics to an external identified database (including, for example, names or Social Security numbers) in a conventional re-identification attack. The Census Bureau attempted to do this but only a small fraction of re-identifications actually turned out to be correct, and Abowd ... concluded that ‘the risk of re-identification is small.’ Therefore, the system worked as designed: because of the combination of swapping, imputation and editing, reporting error in the census, error in the identified credit agency file, and errors introduced in the microdata reconstruction, there is sufficient uncertainty in the data to make positive identification by an outsider impossible.”

This statement may reflect the state of affairs prior to the re-identification ef-

---

fort of the Census discussed in Section 4.1 that succeeded in re-identifying 17% of the US population in 2010. An earlier re-identification attempt by the Census had some issues matching the Census geo-ids with those of commercial data. Once this was understood and fixed, the results discussed in Section 4.1 were obtained.

Ruggles also argues that use of differential privacy will mask respondents characteristics, data that are valuable in demographic and other studies. He correctly asserts that masking characteristics is not explicitly required under the law. But Census is prohibited from publishing

“any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means...”

Given the level of re-identification that was achieved in the Census vulnerability study, it is certainly arguable that releasing tabular information without noise such that the microdata can be reconstructed and possibly matched with external data makes the tabular information just such a representation.

Ruggles further argues that Census would not validate any potential re-identification. This is true, but the fact remains that a commercial data provider can still perform the re-identification attack, then perform a probabilistic record match (perhaps using data held out from the re-identification), and, if the result looks sufficiently promising, add this to their database along with extra information on race, children, sexuality, etc. The argument that Census will not confirm the re-identification is true whether one performs any disclosure avoidance or not. But it is still the responsibility of Census not to abet such re-identification. Finally, there is the issue of whether Census data (as opposed to ACS data) is particularly sensitive. It can be argued that knowledge of various characteristics combined with location data could certainly be abused in various instances and so this provides further support that Census should enforce privacy of census data.

Even more concern has been voiced in the social science and demographer

---

communities regarding the possibility that the ACS tables and microdata sample may also now require similar protection. To date Census has not established that a similar vulnerability exists for the ACS data. Intuitively, it *should* be harder to re-identify this data as it is a small sample of the population and what is released is carefully chosen so as to preserve confidentiality. In any case, no plan by Census exists at present to apply methods of formal privacy to the ACS, and no changes are envisioned in the format for data release at least until 2025 when the issue will be reconsidered (cf. for example, [33]).

## 8.2 Two Statutory Requirements are in Tension in Title 13

It is to be expected that advances in technology may introduce tensions or conflicts among statutory provisions that were seen as conflict-free when they were enacted in the past. Under the Executive Branch’s broad powers to interpret and apply the law, responsibility falls on Executive agency government officials to set policies that attempt to “square the circle” in a defensible manner, even when no perfect solution is possible. Such policies, both as to the procedure of how they are set and their substance, are potentially subject to judicial review, e.g., under the Administrative Procedures Act (5 USC Section 500). The resolution of statutory conflicts is thus ultimately a matter for the courts, or for Congress if it chooses to change the law.

In the above light, we examine two statutory provisions of Title 13. Section 214 (“Wrongful disclosure of information”) provides

“[No official] may make any publication whereby the data furnished by any particular establishment or individual under this title can be identified...”

There is little or no case law to guide us in the interpretation of what, at first sight, seems a clear provision. But how clear is it? Does “whereby” mean by itself

---

without reference to other sources of (e.g., commercial) data? Or does “whereby” mean may not add, even incrementally in the smallest degree, to the likelihood that an individual can be identified using commercially available data? Or is it something in-between? What about “can be identified”? Does this mean identified with certainty? Or does it mean identified probabilistically as more likely than other individuals? And, if the latter, what is the quantitative level of probability that is prohibited?

Census has traditionally adopted very strict interpretations of Section 214 for a host of good reasons, including that doing so encourages trust and participation in the census. Section 141 (Public Law PL 94-171) specifies a process by which the states propose, and the Secretary of Commerce agrees to, a geographical specification of voting districts within each state<sup>3</sup>. It then requires that

“Tabulations of population for the areas identified in any plan approved by the Secretary shall be completed by him as expeditiously as possible after the decennial census date and reported to the Governor of the State involved and to the officers or public bodies having responsibility for legislative apportionment or districting of such State ... ”

The plain-language meaning of “tabulation of population” is fairly obvious: one counts the number of persons satisfying some required condition(s) and enters that number into a table. At the time of the 2010 Census, and with the disclosure avoidance procedures adopted at that time, there seemed to be no significant conflict between the statutory requirements of Section 214 and Section 141. Swapping, for example, preserves population counts in any geographical area. To the extent that swapped individuals were matched for other characteristics (e.g., voting age), counts of persons with matched characteristics would also be preserved. Finally, the use of swapping may allow for the use of a larger value of  $\epsilon$  used for

---

<sup>3</sup>Technically the law says “...the geographic areas for which specific tabulations of population are desired”. This has been identified as blocks and voting districts since the law was passed



---

publication of the various tabulations. This would have to be determined through an empirical assessment of re-identification risk performed both with and without swapping.

Census has determined, and JASON agrees, that swapping alone is an insufficient disclosure avoidance methodology for the 2020 Census. The proposed use of DP in the 2020 Census, which is by now almost certain, will bring the mandates of Section 214 and Section 141 into conflict to a substantially greater degree than previously. Although Census proposes to impose invariants along a backbone of nested geographical regions, the revised state voting districts may not be on this backbone, and hence will be subject to count errors whose magnitude depends on the amount of DP imposed (i.e., the choice of  $\epsilon$ ).

There is no perfect resolution of the conflict. JASON heard the opinion of some experts outside of government that inaccuracies as large as 1000 persons in state voting district counts are acceptable. However, we also heard that, in many cases, the actions of state officials can be interpreted as indicating a mistaken belief that the counts are much more accurate than this. We are not aware of any case law or judicial guidance on the issue. Thus, Census will need to adopt a policy that is a sensible compromise between conflicting provisions of law, recognizing that the ultimate adjudication of such a policy - should it prove to be controversial - lies elsewhere. Too small a value of  $\epsilon$ , while more perfectly satisfying Section 214, satisfies Section 141 less perfectly, both being statutory requirements.

We conclude this report with JASON's findings and recommendations.

## **8.3 Findings**

### **8.3.1 The re-identification vulnerability**

- The Census has demonstrated the re-identification of individuals using the published 2010 census tables.

- 
- Approaches to disclosure avoidance such as swapping and top and bottom coding applied at the level used in the 2010 census are insufficient to prevent re-identification given the ability to perform database reconstruction and the availability of external data.

### **8.3.2 The use of Differential Privacy**

- The proposed use by Census of Differential Privacy to prevent re-identification is promising, but there is as yet no clear picture of how much noise is required to adequately protect census respondents. The appropriate risk assessments have not been performed.
- The Census has not fully identified or prioritized the queries that will be optimized for accuracy under Differential Privacy.
- At some proposed levels of confidentiality protection, and especially for small populations, census block-level data become noisy and lose statistical utility.
- Currently, Differential Privacy implementations do not provide uncertainty estimates for census queries.

As has been seen in Section 6, as the geographic resolution becomes finer, DP will by design affect query results. In such cases, there will at least be a need to inform users of the variances associated with a given query. While the amount of noise injected into tables is known as a result of the open publication of the privacy budgets, the variance in a query is also affected by the size of the population involved in answering that query, the use of the high-dimensional matrix method, the enforcement of invariants, etc. complicating the error analysis. Error assessment could be accomplished by performing multiple instances of a query and then assessing the variation of the results, but this requires re-accessing the data and so potentially violating the DP bounds. Ashmeade [2] has proposed an approach to

---

estimate query error by using the post-processed results and then assessing variance using those results. This has the advantage that one need not access the confidential data. Ashmeade presents some empirical evidence that, for the most part, this approach yields sensible bounds, but for small privacy budgets occasional outliers occur and the results of such an estimate vary widely from the true results obtained using Monte-Carlo methods. This issue clearly requires further work.

### **8.3.3 Stakeholder response**

- Census has not adequately engaged their stakeholder communities regarding the implications of Differential Privacy for confidentiality protection and statistical utility.
- Release of block-level data aggravates the tension between confidentiality protection and data utility.
- Regarding statistical utility, because the use of Differential Privacy is new and state-of-the-art, it is not yet clear to the community of external stakeholders what the overall impact will be.

### **8.3.4 The pace of introduction of Differential Privacy**

- The use of Differential Privacy may bring into conflict two statutory responsibilities of Census, namely reporting of voting district populations and prevention of re-identification.
- The public, and many specialized constituencies, expect from government a measured pace of change, allowing them to adjust to change without excessive dislocation.

---

## 8.4 Recommendations

### 8.4.1 The re-identification vulnerability

- Use substantially equivalent methodologies as employed on the 2010 census data coupled with probabilistic record linkage to assess re-identification risk as a function of the privacy-loss parameter.
- Evaluate the trade-offs between re-identification risk and data utility arising from publishing fewer tables (e.g. none at the block-level) but at larger values of the privacy-loss parameter.

### 8.4.2 Communication with external stakeholders

- Develop and circulate a list of frequently asked questions for the various stakeholder communities.
- Organize a set of workshops wherein users of census data can work with differentially private 2010 census data at various levels of confidentiality protection. Ensure all user communities are represented.
- Develop a set of 2010 tabulations and microdata at differing values of the privacy-loss parameter and make those available to stakeholders so that they can perform relevant queries to assess utility and also provide input into the query optimization process.
- Develop effective communication for groups of stakeholders regarding the impact of Differential Privacy on their uses for census data.
- Develop and provide to users error estimates for queries on data filtered through Differential Privacy.

---

### 8.4.3 Deployment of Differential Privacy for the 2020 census and beyond

- In addition to the use of Differential Privacy, at whatever level of confidentiality protection is ultimately chosen, apply swapping as performed for the 2010 census so that no unexpected weakness of Differential Privacy as applied can result in a 2020 census with less protection than 2010.

There is always the possibility that unforeseen issues or implementation errors may lead to violations of the privacy protections that DP aims to enforce. Such things have happened in the past, for example, in the cryptographic community. JASON recommends that Census apply the traditional disclosure avoidance procedures applied in the 2010 census and then apply DP on top of this dataset. The advantage in JASON's view is that one can communicate that DP is a proposed improvement over traditional approaches and, should there arise any issue with DP, the previously used protections will still be in force. The software infrastructure for the traditional disclosure avoidance approach would have to be reconstructed and this could prove to be a challenge.

- Defer the choice of the privacy-loss parameter and allocation of the detailed privacy budget for the 2020 census until the re-identification risk is assessed and the impact on external users is understood.
- Develop an approach, using real or virtual data enclaves, to facilitate access by trusted users of census data with a larger privacy-loss budget than those released publicly.
- Forgo any public release of block-level data and reallocate that part of the privacy-loss budget to higher geographic levels.
- Amid increasing demands for more granular data and in the face of conflicting statutory requirements, seek clarity on legal obligations for protection of data.

---

## A APPENDIX: Information Theory and Database Uniqueness

Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte.

*(I'd not have made this [letter] so long, had I had time to make it shorter.)*

Blaise Pascal, *Lettres Provinciales*, 4 Dec. 1656.

In this appendix we examine the Dinur-Nissim (DN) results in the context of information theory. As a reminder, DN idealize a database as a string  $d = (d_1, \dots, d_n)$  of  $n$  bits, and a *noiseless* query as the sum of a specified subset of those bits; that is to say, the answer to the query is

$$A(q) = \sum_{i \in q} d_i \equiv \mathbf{w}_q^T \mathbf{d} \tag{A-1}$$

In the second form above, the string  $d$  is represented by a column vector  $\mathbf{d}$ , whose components are either 0 or 1, while  $\mathbf{w}_q^T$  is a row vector of weights applied to the bits before summation; these weights are also 0 or 1, the total number of nonzero weights in  $\mathbf{w}_q$  being denoted  $\#q$ , the size of the subset of bits that this query interrogates. Clearly  $A(q)$  is an integer (a *count*) in the range  $\{0, \dots, \#q\}$ . There are of course  $2^n$  possible distinct queries.

### A.1 Noiseless Reconstruction via Linear Algebra

Each noiseless query constitutes a linear constraint on the  $n$  bits, and distinct queries obviously constitute linearly independent constraints. Here “linear” and “independent” are used in the sense of linear algebra, which therefore guarantees that  $n$  independent queries are *sufficient* to reconstruct  $d$ . Since, however, each component of  $d$  (viewed as a vector in  $\mathbb{R}^n$ ) is restricted to only two possible values, reconstruction may be possible with fewer than  $n$  queries.

---

In what follows, we will often speak of the “probability” of the value of a given bit or bits in the database. In the real world, the noiseless database is fixed, so its bits are not random variables. But in order to be able to apply information-theoretic arguments to the noiseless case, let’s imagine that we are designing a reconstruction algorithm to be applied to the ensemble of *all possible* databases of  $n$  bits. In this ensemble, each bit takes on the values 0 or 1 with equal frequencies ( $= 1/2$ ). To the extent that the actual database can be regarded as having been chosen “at random,” the values of its bits can be regarded as independent random variables.

With this prolog, consider a reconstruction scheme in which we first query  $n/2$  disjoint pairs of bits: e.g., the  $k^{\text{th}}$  query  $q_k$  interrogates bits  $2k - 1$  and  $2k$ , for  $k \in \{1, \dots, n/2\}$ . In the average over all  $2^n$  possible data bases, since each of the two bits interrogated is  $\pm 1$ ,

$$A(q_k) = \begin{cases} 0 & \text{with probability } 1/4, \\ 2 & \text{with probability } 1/4, \\ 1 & \text{with probability } 1/2 \end{cases}$$

When either of the first two possibilities is realized, both bits interrogated by  $q_k$  are determined. Thus we may expect to reconstruct  $n/2$  of the bits with these  $n/2$  queries—a plausible result! But, we now have partial information about the remaining  $n/2$  bits that belong to “ambiguous” pairs where  $A(q_k) = 1$ : namely, the two bits of such a pair must be distinct. There will be approximately  $n/4$  ambiguous pairs. Thus a further  $\sim n/4$  queries that interrogate only the first member of each such pair will resolve the remaining ambiguities. By this argument, we may reconstruct the database with no more than  $\sim 3n/4$  queries. This is fewer than would suffice by the linear-algebra argument, but not by much; which suggests that the linear-algebra argument, though not rigorous, may be useful. As we show in the following subsections, however, it may be possible to do still better—i.e. fewer queries needed for noiseless reconstruction—by a logarithmic factor.

## A.2 Information: An Introductory Example

To further illustrate the point, take the simple case of a 3-bit database. Let  $(B_1, B_2, B_3)$  represent these bits,  $B_i \in \{0, 1\}$ , each with probabilities  $\Pr(B_i = 0) = \Pr(B_i = 1) = \frac{1}{2}$ . Consider two queries,  $Q_L = B_1 + B_2$  (which interrogates the two leftmost bits) and  $Q_R = B_1 + B_3$ . There are of course 8 possible databases, and three possible values for each query, as shown in Table A-4 below:

$B_1$	$B_2$	$B_3$	$Q_L$	$Q_R$
0	0	0	0	0
0	0	1	0	1
0	1	0	1	1
0	1	1	1	2
1	0	0	1	0
1	0	1	1	1
1	1	0	2	1
1	1	1	2	2

Table A-4: Two queries on a 3-bit database

All 8 rows are equally probable. The *entropy* of the joint distribution (probability mass function or PMF) of the three bits is therefore

$$H(B_1, B_2, B_3) = - \sum_{B_1, B_2, B_3} P(B_1, B_2, B_3) \log_2 P(B_1, B_2, B_3) = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3,$$

as one might expect. Notice that in 6 out of 8 cases, the values of the three bits are fully determined by the values of  $(Q_L, Q_R)$ . The exceptions are those in which  $Q_L = Q_R = 1$ , there being two bit combinations 010 and 101 that give this result. So in 3/4 of the cases, two queries suffice to determine the bits, while in the remaining 1/4, a third query is needed. Thus the *average* number of queries needed to reconstruct the database is<sup>4</sup>

$$\frac{3}{4} \times 2 + \frac{1}{4} \times 3 = 2.25 \quad \text{queries on average}$$

<sup>4</sup>One might ask whether it's possible to do better with a different pair of initial queries. There are 28 possible pairs  $[2^3 \times (2^3 - 1)/2]$ , but none does better than this pair.



Another way to look at this is to say that in 3/4 of the cases, the two queries yield 3 bits worth of information; while in the remaining 1/4 of the cases, the queries leave one bit's worth of ambiguity (the choice between databases 010 and 101), so that then in effect they yield only 2 bits of information. Thus the average number of bits of information yielded by these two queries is

$$\frac{3}{4} \times 3 + \frac{1}{4} \times 2 = 2.75 \quad \text{bits of information on average}$$

The joint PMF of  $(Q_L, Q_R)$ , which follows from Table A-4, is

$Q_L$	$Q_R$	probability
0	0	1/8
0	1	1/8
1	0	1/8
1	1	2/8
0	2	0
2	0	0
1	2	1/8
2	1	1/8
2	2	1/8

Table A-5: Joint probability mass function of two queries.

The entropy of these two variables is therefore (combinations that have zero probability being omitted from the sum)

$$-\sum_{Q_L, Q_R} P(Q_L, Q_R) \log_2 P(Q_L, Q_R) = -6 \times \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{4} \log_2 \frac{1}{4} = 2.75$$

Evidently, the entropy of the PMF of  $(Q_L, Q_R)$  coincides with the average number of bits of information gained from these two queries. This generalizes.

Looking ahead to Section A.4, the covariance of these two queries is

$$C = \text{cov}(Q_L, Q_R) = \frac{1}{4} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

and the Gaussian approximation described there predicts that

$$H(Q_L, Q_R) \approx \frac{1}{2} \log_2 \det(2\pi e C) \approx 2.88667$$

This is an overestimate (2.88667 instead of 2.75), presumably because the Gaussian approximation is not accurate for queries involving small numbers of bits. Yet it is qualitatively correct: 2 well-chosen queries on 3 bits yield  $> 2$  but  $< 3$  bits of information on average.

### A.3 Information Gained Per Query

In the examples above, why do we do better by querying two bits at a time, and how can this be generalized?

Querying a single bit—noiselessly—reaps exactly one bit of information, because there are two possible outcomes (0 or 1), and averaged over all possible databases, these outcomes have equal frequency.

Consider now a query  $q$  that sums  $\#q = m \geq 1$  bits. There are now  $m + 1$  possible values for the answer  $A(q) = a \in \{0, \dots, m\}$ . In the data-base ensemble, the probabilities or frequencies  $\{f_a\}$  of these outcomes have the binomial distribution  $B(m, 1/2)$ , meaning that

$$f_a = 2^{-m} \binom{m}{a}, \quad \Rightarrow \quad \sum_a f_a = 1. \quad (\text{A-2})$$

The formal information gained from this query is then

$$I(A) = - \sum_a f_a \log_2 f_a \quad (\text{A-3a})$$

$$\approx \frac{1}{2} \log_2 m + \underbrace{\frac{1}{2} \log_2(\pi e/2)}_{\approx 1.047096} \equiv I_G(A) \quad (\text{A-3b})$$

The second line is obtained by approximating the binomial distribution as a Gaussian (with mean  $E(A) = m/2$  and variance  $m/4$ ). Table A-6 shows that the Gaussian approximation is quite good even for small  $m$ —but not for  $m = 0$ , a point that will be important in Section A.7.

---

$m$	$I$	$I_G$
0	0	$-\infty$
1	1	1.047096
2	3/2	1.547096
16	3.04655	3.047096
128	4.547088	4.547096

Table A-6: Average information gain, in bits, from a single noiseless query that sums  $m$  bits. Second column is exact; third column is the Gaussian approximation.

What we have called  $I(m)$  is also the *entropy*  $H(X)$  of a binomially distributed random variable  $X \sim B(m, 1/2)$ . We use the notation  $I$  rather than  $H$  in this instance because we think of it as measuring the average *knowledge gained* after a query, rather than the *uncertainty* in the outcome of the query. But regardless of the interpretation, the mathematical rules governing information/entropy are the same.

## A.4 Information Gained from Multiple Noiseless Queries

The preceding discussion shows that the most informative *single* query is the sum of all  $n$  bits: the information gained is  $I(n) \approx 0.5 \log_2(n)$  for  $n \gg 1$ . But of course this is not enough to reconstruct all  $n \gg \log_2 n$  bits. Clearly reconstruction requires multiple queries; but what is the minimum number? One may speculate that since a single query  $q$  that sums  $\#q \sim O(n)$  bits yields  $O(\log n)$  bits of information, it should follow that the minimum number of such queries required is  $O(n/\log n)$ . But this is not obvious, because queries are not independent unless they interrogate disjoint subsets of the  $n$  bits. Therefore their information will not simply add. In the first two schemes above, the subsets *were* independent: those queries interrogated individual bits or disjoint pairs of bits. But such “small” queries [ $\#q \sim O(1)$ ] yield less information (at least individually) than “large” queries [ $\#q \gg 1$ ]. And for  $n \gg 1$ , since we will need *at least*  $O(n/\log n)$  queries to reconstruct, they cannot be entirely disjoint if they are individually

large.

Consider now two queries  $q_1$  and  $q_2$ , and let  $q_1 \cap q_2$  be the subset of bits that they have in common. If these queries are large, i.e.,  $\min(\#q_1, \#q_2) \gg 1$ , then by the Central Limit Theorem, they are well approximated as Gaussian random variables, with means  $E(q_i) = \frac{1}{2}\#q_i$  for  $i \in \{1, 2\}$ , and covariance matrix

$$\mathbf{C} = \frac{1}{4} \begin{pmatrix} \#q_1 & \#(q_1 \cap q_2) \\ \#(q_1 \cap q_2) & \#q_2 \end{pmatrix}$$

(The prefactor comes from the fact that the mean-subtracted bit values are  $\pm\frac{1}{2}$ , whence the variance of individual bits is  $\frac{1}{4}$ .) It is easily seen that if the “information” of a multivariate Gaussian density function

$$P(\mathbf{x})d\mathbf{x} = \frac{1}{\sqrt{\det(2\pi\mathbf{C})}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x}\right) d\mathbf{x}$$

is defined by  $-\int P(\mathbf{x}) \log_2 P(\mathbf{x}) d\mathbf{x}$ , then this information is

$$I(\mathbf{C}) = \log_2 \sqrt{\det(2\pi e\mathbf{C})}, \quad (\text{A-4})$$

This reduces to the Gaussian approximation of Section A.3 for a single query, where  $\mathbf{C} \rightarrow m/4$ , a scalar. For multiple *disjoint* queries, so that  $\mathbf{C}$  is diagonal, eq. (A-4) says that the total information is the sum of the informations gained from each query separately. If the queries are not disjoint, then at least some of the off-diagonal entries of  $\mathbf{C}$  are positive, and none are negative, whence the determinant of  $\mathbf{C}$  is less than the product of its diagonals: this means that the total information is less than the sum of the information obtained from the individual queries.

The goal now is to find the smallest rank  $r$  (i.e., the smallest number of queries) for which  $I(\mathbf{C}) > n$ , with the restriction that

$$\mathbf{C} = \frac{1}{4}\mathbf{W}^T\mathbf{W}, \quad (\text{A-5})$$

for some  $n \times r$  matrix  $\mathbf{W}$  whose entries are 0 or 1: each column of  $\mathbf{W}$  corresponds to a query vector  $\mathbf{w}_q$ . If the information  $I(\mathbf{C}) > n$ , we can expect to be able to reconstruct “most”  $n$ -bit databases with these  $r$  queries.

---

Suppose, to begin with, that the entries of the matrix  $\mathbf{W}$  are chosen at random. In this case, approximately half of the elements in each column (i.e., in each query vector) would be 1, and the remainder 0; but the excess or deficit of 1s over 0s in each column would fluctuate by  $O(\sqrt{n})$ . Any two distinct columns of  $\mathbf{W}$  would have approximately  $n/4$  1s in common, so that  $\sum_k W_{ik}W_{kj} \approx (n/4)(1 + \delta_{ij})$ . The elements of the covariance matrix would then be

$$C_{ij} = \begin{cases} n/8 + O(\sqrt{n}) & \text{if } i = j \in \{1, \dots, r\} \\ n/16 + O(\sqrt{n}) & \text{if } i \neq j \end{cases} \quad (\text{A-6})$$

The  $O(\sqrt{n})$  are random in sign and have mean 0, so that it might be hoped that in computing  $\log_2 \det \mathbf{C}$  for sufficiently large  $n$ , we could neglect them compared to the  $O(n)$  terms. The matrix with these terms neglected is

$$\bar{\mathbf{C}} = \frac{n}{16} (\mathbf{I}_r + \mathbf{J}_r), \quad (\text{A-7})$$

in which  $\mathbf{I}_r$  is the  $r \times r$  identity matrix, and the matrix  $\mathbf{J}_r$  is entirely filled with 1s (sometimes called the “unit” matrix, although this risks confusion with the identity). Since  $\mathbf{I}_r$  commutes with  $\mathbf{J}_r$ , the two matrices can be simultaneously diagonalized, and their eigenvalues simply add.

It is not hard to see that the eigenvectors of  $\mathbf{J}$  have the form

$$\mathbf{v}_\omega = (1, \omega, \omega^2, \dots, \omega^{r-1})^T$$

with  $\omega^r = 1$ , i.e.  $\omega$  is any of the  $r^{\text{th}}$  roots of unity. These eigenvectors are orthogonal ( $\mathbf{v}_\omega^\dagger \mathbf{v}_{\omega'} = r\delta_{\omega, \omega'}$ ), as is familiar from the Discrete Fourier Transform. For the trivial root  $\omega = 1$ , the eigenvalue of  $\mathbf{J}$  is  $r$ , while all of the  $r - 1$  other roots correspond to zero eigenvalues. Therefore the eigenvalues of  $\mathbf{I} + \mathbf{J}$  are

$$\underbrace{\{1, \dots, 1\}}_{r-1 \text{ times}}, 1 + r\}$$

and it follows that

$$\begin{aligned} I(\bar{\mathbf{C}}) &\equiv \frac{1}{2} \log_2 \det(2\pi e \bar{\mathbf{C}}) \\ &= \frac{1}{2} r \log_2 \left( \frac{\pi e}{8} n \right) + \frac{1}{2} \log_2(1 + r) \\ &\approx \frac{1}{2} r (\log_2 n + 0.094) \quad \text{for } r, n \gg 1. \end{aligned} \quad (\text{A-8})$$

---

## A.5 $m$ Sequences and Hadamard Matrices

The replacement

$$C \rightarrow \bar{C}$$

is an approximation. But we can obtain the determinant (A-7) exactly in the special cases that  $n = 2^k - 1$  through a cunning *pseudorandom* choice of the query vectors: namely,  $m$ -sequences, a.k.a. maximum-length Linear Feedback Shift Register (LFSR) sequences [11]. In the form we need them here, they are periodic sequences of bits  $b_i \in \{0, 1\}$  with period  $n = 2^k - 1$  and autocorrelation function

$$A(j) \equiv \sum_{i=0}^{n-1} b_i b_{i+j} = \begin{cases} (n+1)/2 & \text{when } j \equiv 0 \pmod{n} \\ (n+1)/4 & \text{otherwise} \end{cases} \quad (\text{A-9})$$

If we populate the columns of  $\mathbf{W}$  with distinct circular shifts of such a sequence, then  $\mathbf{C}$  will have almost exactly the form (A-7), the only change being that  $n \rightarrow n+1$  (an even number). Then the information gained from these  $r$  queries will be exactly as in the second line of (A-8), except for the same replacement.<sup>5</sup>

Hadamard matrices yield similarly good correlation properties [11]. By definition, a Hadamard matrix of order  $n$  is an  $n \times n$  matrix  $\mathbf{H}$  whose entries are  $\pm 1$  and whose rows are orthogonal, so that  $\mathbf{H}\mathbf{H}^T = n\mathbf{I}$ , where  $\mathbf{I}$  is the  $n \times n$  identity. The order  $n$  must be 1, 2, or a multiple of 4; it is conjectured but not proved that Hadamard matrices exist for every multiple of 4. There are explicit constructions for special cases, however, and in particular for  $n = p + 1$  where  $p$  is a prime of the form  $4k - 1$  (i.e.  $n \in \{4, 8, 12, 20, 24, 32, 44, 48, 60, \dots\}$ ). Importantly, the first row (and first column) of the latter sort<sup>6</sup> of Hadamard matrix is all 1s, so it follows from the definition that each of the remaining rows has an equal number of  $+1$ s and  $-1$ s. It is then not hard to see that if we replace the elements  $H_{ij}$  of such a matrix with

$$W_{ij} = \frac{1}{2}(H_{\sigma(j)i} + 1),$$

---

<sup>5</sup>Exact, that is, within our Gaussian approximation for the binomial query outcomes.

<sup>6</sup>a “cyclic” Hadamard matrix [11]

so that the  $j^{\text{th}}$  column of  $\mathbf{W}$  is the  $\sigma(j)^{\text{th}}$  row of  $\mathbf{H}$  with every  $-1$  replaced by  $0$ , then the elements of  $\mathbf{W}^T \mathbf{W}$  are

$$\sum_{i=1}^n W_{ij} W_{ik} = \begin{cases} n & j = k \text{ \& } \sigma(j) = 1, \\ n/2 & j = k \text{ \& } \sigma(j) \neq 1, \\ n/2 & j \neq k \text{ \& } \min(\sigma(j), \sigma(k)) = 1, \\ n/4 & j \neq k \text{ \& } \min(\sigma(j), \sigma(k)) \neq 1. \end{cases} \quad (\text{A-10})$$

Here  $\sigma()$  is any permutation of  $\{1, 2, \dots, n\}$ . But we do not have to use the complete permutation: we can use a part of it that selects some subset of  $r$  rows from  $\mathbf{H}$ , in which case  $\mathbf{W}$  becomes  $n \times r$ , while the covariance matrix  $\mathbf{C} \equiv \frac{1}{4} \mathbf{W}^T \mathbf{W}$  becomes  $r \times r$ . If this subset does not include the first row of  $\mathbf{H}$  (the row that is all 1s), then  $\mathbf{C}$  has exactly the form (A-7), and hence the same eigenvalues and determinant. If the first row of  $\mathbf{H}$  is included, then the eigenvalues and determinant can be found by Cholesky decomposition  $\mathbf{C} = \mathbf{L} \mathbf{L}^T$ , where  $\mathbf{L}$  is lower triangular.

The diagonal entries of  $\mathbf{L}$  are the square roots of the eigenvalues of  $\mathbf{C}$ . It turns out that when the first column of  $\mathbf{W}$  is the first row of  $\mathbf{H}$ , the first diagonal of  $\mathbf{L}$  is  $\sqrt{n}/2$ , all the rest are  $\sqrt{n}/4$ , and the rest of  $\mathbf{L}$  vanishes except for the first column, in which all the elements after the first are also  $\sqrt{n}/4$ . In this case, all of the eigenvalues of  $\mathbf{C}$  coincide with those of (A-7) (i.e., they are  $n/16$ ) except for the first, which is  $n/4$  in this case, but  $n(r+1)/16$  in (A-7). So if  $r < n$  (fewer queries than bits), it is slightly advantageous not to use the first row of  $\mathbf{H}$ , i.e. not to include the query that sums all of the bits.

## A.6 The Minimal Number of Queries

We have seen that, within our Gaussian approximation at least, and neglecting  $O(1)$  corrections, the information gained from  $r \leq n$  noiseless queries on an  $n$ -bit database can be made as large as

$$\max(I_r) \approx \frac{r}{2} [\log_2 n + \log_2(\pi e/8)].$$

On the other hand, it follows from eq. (A-3a) that the maximum information obtained from a single query is  $\max(I_1) \lesssim \log_2 n + \log_2(\pi e/2)$ : we do best by sum-

---

ming all of the  $n$  bits. It would seem therefore that the redundancy among multiple queries can be made almost negligible, i.e.  $\max(I_r) \approx r \max(I_1)$ : the information contributed by distinct queries is almost additive, apart from the different constants  $\log_2(\pi e/2)$  vs.  $\log_2(\pi e/8)$ .

In the absence of prior constraints on the bits in the database, we must have  $I_r \geq n$  in order to determine all of the bits. Thus

*The minimum number of noiseless queries needed to reconstruct an  $n$ -bit database is at least  $2n/\log_2 n$  for large  $n$ .*

We have tested this by numerical experiments with modest values of  $n$  and  $r$ , as shown in Table A-7. Using a modified hill-climbing technique, we have constructed a set of near-optimal (better than random) queries<sup>7</sup>. As shown in the fourth column, most of the  $2^n$  possible databases answer our  $\lceil 2n/\log_2 n \rceil$  queries uniquely, but not all. As we add queries, the number of ambiguous cases appears to drop exponentially. The third column shows the minimum number of queries needed to resolve all ambiguities. The evidence of this table suggests that the  $r \sim 2n/\log_2 n$  criterion is relevant, but because exhaustion over all  $2^n$  databases is impractical for much larger  $n$ , it is also consistent with the possibility that the minimum  $r/n$  needed to resolve all ambiguities asymptotes to a constant. This is what was found empirically in Section 4 but it's important to note that there is no guarantee that the least squares approach used there is optimal in the Shannon or information-theoretic sense.

## A.7 Noisy Single Queries

Instead of the exact answer (A-1) to a query, we receive a noisy version  $\hat{A}(q) = \mathbf{w}_q^T \mathbf{d} + N_q$ , where  $N_q$  is a random variable independent of the database and query

---

<sup>7</sup>by attempting to maximize  $\mathbf{W}^T \mathbf{W}$ , with the restriction that  $\mathbf{W}$  is  $n \times r$  and its entries are all 0 or 1



---

$n$	$\lceil 2n/\log_2 n \rceil$	$r_{\min}$	uniques
8	6	6	98.4%
9	6	6	100%
10	7	7	100%
11	7	8	96.9%
12	7	9	88.7%
13	8	9	96.1%
14	8	9	94.6%
15	8	9	90.1%
16	8	10	83.5%
17	9	11	93.8%
18	9	13	88.0%
19	9	13	79.3%
20	10	14	95.8%
21	10	14	90.9%

Table A-7: Numerical experiments on noiseless queries of small databases. 2nd column is the smallest integer  $\geq 2n/\log_2 n$ . 3rd column is the minimum number of optimized queries needed to determine all  $2^n$  databases uniquely. 4th is the fraction that are uniquely identified by  $\lceil 2n/\log_2 n \rceil$  queries.

vectors. For convenience, the noise variables  $N_q$  and  $N_{q'}$  belonging to distinct queries  $q$  and  $q'$  will be assumed independent and identically distributed.<sup>8</sup>

Presumably also there is a rule that a given query can be asked at most once—or if not, that the value taken by  $N_q$  is the same every time that query is asked: for if not, it would be possible to beat down the noise by asking the query repeatedly and averaging the answers.

The concept of *mutual information*  $I(X, Y)$  is useful to express the knowledge that one has of a random variable  $X$  given an observation of a second variable  $Y$ , which for this application is a noisy version of  $X$  (Fig. A-1).

---

<sup>8</sup>This is not essential. In fact, the High Dimensional Matrix Method used by Census [19] creates correlations among the  $N_q$ . As long as the noise remains independent of the database, the effect is to replace the noise covariance matrix  $\sigma_N^2 \mathbf{I}$  in eq. (A-14) with some other (symmetric) matrix.

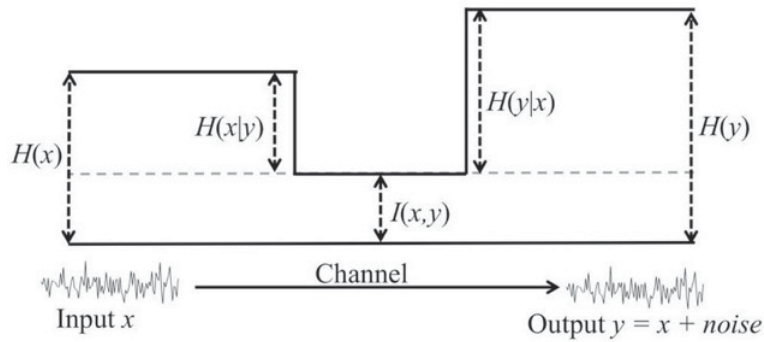


Figure A-1: Communication over a noisy channel.  $X$  ranges over transmitted signals, and  $Y$  over the noisy versions received. The entropy  $H(X)$  is the minimum number of noiseless bits required to specify the value of  $X$ , and similarly for  $H(Y)$ .  $H(X|Y)$  is the average uncertainty ( $\sim$ unknown bits) in  $X$  given a measurement of  $Y$ . The difference  $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$  is the mutual information.

The formal definition for discrete variables is

$$I(X;Y) = \sum_{X=x} \sum_{Y=y} p_{X,Y}(x,y) \log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}. \quad (\text{A-11})$$

Here the sums are taken over all possible values  $x$  and  $y$  of  $X$  and  $Y$  respectively, while  $p_X$ ,  $p_Y$ , and  $p_{X,Y}$  are the probability mass functions (PMFs) for  $X$  alone, for  $Y$  alone, and for  $(X,Y)$  jointly. It can be shown that  $I(X;Y) \geq 0$ , with equality iff  $X$  and  $Y$  are independent.

A small example may increase confidence in this definition. Suppose  $X$  represents a single-bit message with equally frequent values  $\{0, 1\}$ , and  $Y = X + N$  with  $N$  a noise bit that is also equally likely to be 0 or 1. Therefore  $Y \in \{0, 1, 2\}$ . The PMFs are described by the following table:

---

$x$	$y$	$p_X(x)$	$p_Y(y)$	$p_{X,Y}(x,y)$
0	0	1/2	1/4	1/4
0	1	1/2	1/2	1/4
0	2	1/2	1/4	0
1	0	1/2	1/4	0
1	1	1/2	1/2	1/4
1	2	1/2	1/4	1/4

The third and fourth entries in the last column (for the joint PMF) vanish, because for example if  $X = 0$  then  $Y = 2$  is impossible, as the noise bit is at most 1. If  $Y = 0$  or  $Y = 2$ , then  $X$  is determined (as 0 or 1, respectively). Taken together, these outcomes happen half the time:  $p_{X,Y}(0,0) + p_{X,Y}(1,2) = 1/2$ . In case  $Y = 1$ , however,  $X$  is equally likely to be 0 or 1. So observing  $Y$  yields perfect knowledge of  $X$  half the time, and the rest of the time no information at all. We may therefore say that observing  $Y$  is worth half a bit of knowledge about  $X$  on average. If one works through the definition (A-11) using the values in this table,<sup>9</sup> one finds indeed that  $I(X;Y) = 1/2$ .

A general theorem about mutual information is[22]

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),$$

in which  $H(X)$  and  $H(Y)$  are the entropies<sup>10</sup> of  $X$  and  $Y$  separately, while  $H(X|Y)$  is the residual entropy of  $X$  after  $Y$  is observed, and similarly for  $H(Y|X)$ . This is illustrated in Fig. A-1. It is easily seen that if  $X$  and  $N$  are independent, then  $H(X + N|X) = H(N)$ . Therefore,

$$I(X; X + N) = H(X + N) - H(N) \quad \text{when } X \text{ is independent of } N. \quad (\text{A-12})$$

Suppose for example that  $X$  and  $N$  are independent univariate Gaussian variables, so that  $Y = X + N$  is also Gaussian, and  $\text{var}Y = \text{var}X + \text{var}N$ . Since the

<sup>9</sup>It is understood that  $0 \cdot \log_2 0 = 0$ , i.e. cases for which  $p_{X,Y}(x,y) = 0$  are excluded from the sum.

<sup>10</sup>See the discussion of entropy vs. information in Section A.3

---

entropy of a Gaussian is<sup>11</sup>  $H(X) = \frac{1}{2} \log_2(2\pi e \text{var} X)$ , and similarly for  $H(Y)$  and  $H(N)$ , it follows that

$$I(X;Y) = \frac{1}{2} \log_2 \left( 1 + \frac{\text{var}(X)}{\text{var}(N)} \right). \quad (\text{A-13})$$

The logarithm here is strongly reminiscent of the factor  $\log_2 \left( 1 + \frac{P_{\text{signal}}}{P_{\text{noise}}} \right)$  in Shannon's channel capacity theorem [31].

To relate this result to the previous discussion of noiseless queries, we need to understand what happens as the variance of the noise tends to zero. In this limit, the Gaussian approximation breaks down. The exact query results ( $X$ ) are actually integers with a binomial distribution. If noise with  $\text{var}(N) \ll 1$  is added to such queries, the exact result ( $X$ ) can be obtained from  $X + N$  by rounding to the nearest integer with negligible probability of error. So we should expect  $I(X, X + N)$  to reduce to  $H(X)$ , which is finite, as  $\text{var}(N) \rightarrow 0$ . However, eq. (A-13) presumes that both  $X$  and  $N$  take real values, and it yields an infinite result as  $\text{var}(N) \rightarrow 0$  because arbitrarily close real numbers can always be distinguished.

Suppose instead that both  $X$  and  $N$  are discrete independent variables, for example with binomial distributions  $B(m, 1/2)$  and  $B(m', 1/2)$  respectively. Then  $Y = X + N$  is distributed as  $B(m + m', 1/2)$ . Also<sup>12</sup>  $\text{var}(X) = m/4$ ,  $\text{var}(N) = m'/4$ , and  $\text{var}(Y) = (m + m')/4$ . If  $m' \geq 1$ , then the Gaussian approximations for  $H(N)$  and  $H(Y)$  are quite accurate, as shown by Table (A-6), so that eq. (A-13) is a good approximation to the mutual information. But in the noiseless case  $m' = 0$ , we have to use the exact definition in the first line of eq. (A-3a) for the entropy of a binomial; this yields  $H(N) = 0$ . Then it follows from eq. (A-12) that  $I(X;Y) \rightarrow I(X;X) = H(X)$ , as we expect, rather than  $+\infty$  as the Gaussian approximation (A-13) would predict in the noiseless limit.

---

<sup>11</sup>For a multivariate Gaussian, this becomes  $H(\mathbf{X}) = \frac{1}{2} \log_2 \det[2\pi e \text{cov}(\mathbf{X})]$ , where  $\text{cov}(\mathbf{X})$  is the covariance matrix of  $\mathbf{X}$

<sup>12</sup>Recall that if  $X \sim B(n, p)$ , where  $p$  is the probability of "success" on a single trial and  $n$  is the number of trials, that  $\text{var}(X) = np(1 - p)$ .

---

## A.8 Multiple Noisy Queries

This generalizes directly to multiple queries, represented by a vector  $\mathbf{X}$  when exact, but corrupted by a noise vector  $\mathbf{N}$  with diagonal covariance  $\text{cov}(\mathbf{N}) = \sigma_N^2 \mathbf{I}$ . Provided  $\sigma_N^2 \gtrsim 1/4$ , we may use the Gaussian approximation, so that

$$I(\mathbf{X}, \mathbf{X} + \mathbf{N}) \approx \frac{1}{2} \log_2 \det[\sigma_N^{-2} \mathbf{C} + \mathbf{I}]. \quad (\text{A-14})$$

in which  $\mathbf{C} = \text{cov}(\mathbf{X})$  is determined as before by the  $n \times r$  query matrix  $\mathbf{W}$  [eq. (A-5)], and  $\mathbf{I}$  is the  $r \times r$  identity.

The result (A-14) should be interpreted as the total information gathered by these queries in the presence of noise. As we've seen in Section A.4, for sensible (e.g. random) choices of the query matrix  $\mathbf{W}$ , all but one of the eigenvalues of  $\mathbf{C}$  is approximately equal to  $n/16$  if  $n \geq r \gg 1$ . It follows that the net information gathered on average is

$$I_{\text{net}} \approx \frac{r-1}{2} \log_2 \left( 1 + \frac{n}{16\sigma_N^2} \right) + \frac{1}{2} \log_2 \left( 1 + \frac{n(r+1)}{16\sigma_N^2} \right). \quad (\text{A-15})$$

(The second logarithm comes from the one nonzero eigenvalue of the matrix  $\mathbf{J}$  discussed above.) If there is to be hope of reconstructing the database, the information  $I_{\text{net}}$  must be  $\geq n$ , the number of bits to be reconstructed. If the standard deviation of the noise  $\sigma_N > \sqrt{n/48}$ , however, then the logarithm  $< 2$ , in which case we will not have enough information even at  $r = n$ —i.e., even if we make as many queries as bits. This is reminiscent of DN's result to the effect that  $O(\sqrt{n})$  noise is sufficient to prevent an “algebraically bounded” adversary from reconstructing the database.

But now suppose that we are allowed to make  $r \gg n$  queries. This is most interesting in the large-noise limit, i.e. where  $\sigma_N^2$  is large compared to all of the eigenvalues of  $\mathbf{C}$ . Note by the way that  $\mathbf{C}$  becomes singular for  $r > n$ , because it is constructed from  $\mathbf{W}$ , which has rank  $\min(r, n)$ . However, the combination  $\sigma_N^2 \mathbf{C} + \mathbf{I}$  is nonsingular, and for sufficiently large  $\sigma_N^2$ , the expansion

$$\log_e \det(\mathbf{I} + \varepsilon \mathbf{M}) \rightarrow \varepsilon \text{Trace}(\mathbf{M}) + O(\varepsilon^2) \quad \text{as } \varepsilon \rightarrow 0 \text{ at fixed } \mathbf{M}$$

---

allows us to write

$$I_{\text{net}} \approx \frac{\log_2 e}{2\sigma_N^2} \text{Trace}(\mathbf{C}) \approx \frac{nr \log_2 e}{16\sigma_N^2} \quad (\sigma_N^2 \gg n/16) \quad (\text{A-16})$$

Hence, even if the signal-to-noise ratio per query is very small, a sufficient number of queries—specifically,  $r \gtrsim 16\sigma_N^2 / \log_e 2$ —should gather enough information to determine the database. We have not checked this prediction experimentally but we do confirm that it is possible to gather sufficient information to reconstruct the DN database provided we can issue enough queries. Note that this result indicates one will always recover the bits if the variance of the noise is held fixed as the queries are issued.

## A.9 Reconstruction

So far we’ve talked about gathering enough information, through queries, to *determine* the bits in a database; but we haven’t provided a method for actually estimating the bits from the query results. Methods based on bounded least squares optimization are discussed elsewhere in this report, and illustrated by numerical experiments. Here we provide an alternative approach, straightforwardly applying Bayesian inference to our Gaussian approximation. For simplicity, we discuss here only the noiseless case, but the method is easily generalized to include noise.

The general idea is this. We choose a full  $n \times n$  matrix  $\mathbf{W}$  of query weights, with  $\det \mathbf{W}$  nonzero. We then ask, after the first  $r < n$  of these queries (defined by the first  $r$  columns of  $\mathbf{W}$ ) have been posed and answered, what is the posterior (conditional) probability distribution for the answers to the remaining  $n - r$  queries that have not yet been made? If this posterior is narrow, the likely answers to the not-yet-asked queries can be predicted with probable errors less than unity (i.e., less than a bit). Then, from the results of only the first  $r$  queries, we may write down a shrewd estimate for the full  $n \times n$  linear system discussed in Section A.1 and invert for the bits (rounding the real-valued answers to 0 or 1 as needed). If on the other hand the posterior is not narrow enough, we increase  $r$  (i.e., ask more

queries) until it is.

This procedure is in principle well-defined if the queries are treated exactly as discrete binomial variables. But unfortunately we do not know how to make the exact calculations except by brute force. So we resort to our Gaussian approximation. Let  $\mathbf{X}_n$  be the full length- $n$  vector of random variables for the outcomes of all  $n$  queries defined by some  $n \times n$  weight matrix  $\mathbf{W}_n$  with entries  $\in \{0, 1\}$  and  $\det \mathbf{W}_n \neq 0$ . In the Gaussian approximation, the joint distribution of  $\mathbf{X}_n$  is determined by the means  $\boldsymbol{\mu}_n = E(\mathbf{X}_n)$  and covariances  $\mathbf{C}_n = E[\mathbf{X}_n - \boldsymbol{\mu}_n](\mathbf{X}_n - \boldsymbol{\mu}_n)^T$ . As in Section A.4, since we assume uniform priors on all of the database bits (0 or 1 with equal probability), each component of  $m\boldsymbol{\mu}_n$  equals one half the sum of the corresponding column of  $\mathbf{W}_n$ , while  $\mathbf{C}_n = \frac{1}{4}\mathbf{W}_n^T\mathbf{W}_n$ .

Now partition  $\mathbf{X}_n$  into its first  $r$  components  $\mathbf{X}_r$  and the remaining  $n - r$  components  $\mathbf{X}_{n-r}$ , with corresponding partitions of the means and covariances:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_r \\ \boldsymbol{\mu}_{n-r} \end{bmatrix}, \quad \mathbf{C}_n = \begin{bmatrix} \mathbf{C}_r & \mathbf{C}_{r,n-r} \\ \mathbf{C}_{n-r,r} & \mathbf{C}_{n-r} \end{bmatrix} \quad (\text{A-17})$$

Here

$$\mathbf{C}_r = E(\mathbf{X}_r\mathbf{X}_r^T)$$

represents the  $r \times r$  covariances of the components of  $\mathbf{X}_r$  among themselves, and similarly for

$$\mathbf{C}_{n-r} = E(\mathbf{X}_{n-r}\mathbf{X}_{n-r}^T);$$

while

$$\mathbf{C}_{r,n-r} = E(\mathbf{X}_r\mathbf{X}_{n-r}^T)$$

and its transpose

$$\mathbf{C}_{n-r,r} = E(\mathbf{X}_{n-r}\mathbf{X}_r^T)$$

encode the  $r \times (n - r)$  cross-correlations between the components of  $\mathbf{X}_r$  and  $\mathbf{X}_{n-r}$ . As is well known,<sup>13</sup> the conditional probability  $\Pr(\mathbf{X}_{n-r}|\mathbf{X}_r = \mathbf{x}_r)$  is itself

<sup>13</sup>see, e.g., the Wikipedia article ‘‘Multivariate normal distribution’’ and references therein

---

Gaussian, with means and covariances

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{n-r} &= \boldsymbol{\mu}_{n-r} + \mathbf{C}_{n-r,r} \mathbf{C}_r^{-1} (\mathbf{x}_r - \boldsymbol{\mu}_{n-r}) \\ \hat{\mathbf{C}}_{n-r} &= \mathbf{C}_{n-r} - \underbrace{\mathbf{C}_{n-r,r} \mathbf{C}_r^{-1} \mathbf{C}_{n-r,r}^T}_{\mathbf{Q}}.\end{aligned}\quad (\text{A-18})$$

Since the matrix  $\mathbf{Q}$  is positive semidefinite, it follows that  $\det \hat{\mathbf{C}} \leq \det \mathbf{C}_{n-r}$ , with equality only if the cross correlations  $\mathbf{C}_{n-r,r}$  vanish.

Importantly, the reduced covariance matrix  $\hat{\mathbf{C}}$  for the unposed  $n-r$  queries does not depend on the results ( $\mathbf{X}_r = \mathbf{x}_r$ ) of the first  $r$  queries, so we can work it out in advance in terms of the query weights  $\mathbf{W}_n$ . This can be done explicitly when  $\mathbf{C}_n$  has the simple form (A-7), which we can obtain by choosing the columns of  $\mathbf{W}$  to be  $m$  sequences, or by choosing them at random and neglecting the resulting  $O(\sqrt{n})$  “fluctuations” in the resulting components of  $\mathbf{C}$  [eq. (A-8)]. In this case,  $\mathbf{C}_r$  and  $\mathbf{C}_{n-r}$  have similar forms, except that in each case,  $\mathbf{I}$  and  $\mathbf{J}$  are matrices of the appropriate order.<sup>14</sup> It’s clear that  $\mathbf{J}_k^2 = k\mathbf{J}_k$  for every  $k$ , and therefore

$$(\mathbf{I}_k + \mathbf{J}_k)^{-1} = \mathbf{I}_k - \frac{1}{k+1} \mathbf{J}_k$$

The off-diagonal matrix  $\mathbf{C}_{r,n-r} = \frac{n}{16} \mathbf{J}_{r,n-r}$ ,  $\mathbf{J}_{k,m}$  being the  $k \times m$  matrix with all entries equal to 1 (so that  $\mathbf{J}_{k,k} = \mathbf{J}_k$ ). By means of the rules

$$\mathbf{J}_{j,k} \mathbf{I}_k = \mathbf{J}_{j,k} \quad \text{and} \quad \mathbf{J}_{i,k} \mathbf{J}_{k,j} = k \mathbf{J}_{i,j}$$

we can now evaluate the reduced covariance (A-18) for this choice of queries:

$$\hat{\mathbf{C}}_{n-r} = \frac{n}{16} \left( \mathbf{I}_{n-r} + \frac{1}{r+1} \mathbf{J}_{n-r} \right).\quad (\text{A-19})$$

The determinant of  $\hat{\mathbf{C}}_{n-r}$  is smaller than that of  $\mathbf{C}_{n-r} = \frac{n}{16} (\mathbf{I}_{n-r} + \mathbf{J}_{n-r})$  by a factor  $(2r+1)/(r+1)^2 \approx 2r^{-1}$  for  $r \gg 1$ . In logarithmic terms, this is a disappointingly slight reduction in uncertainty.

---

<sup>14</sup>I.e.,  $\mathbf{C}_k = \frac{n}{16} (\mathbf{I}_k + \mathbf{J}_k)$ , with  $\mathbf{I}_k$  being the  $k \times k$  identity, and  $\mathbf{J}_k$  being the  $k \times k$  matrix with all elements equal to 1. The prefactor  $\frac{n}{16}$  in  $\mathbf{C}_k$ , however, is invariant.





---

## B MATLAB CODE FOR DN DATABASE RECONSTRUCTION

The MATLAB codes in this appendix can be used to generate the various figures in the report associated with the calculations on the Dinur-Nissim database.

Listing 1: Matlab script for Figure 5-1

```
1 % script to recover the bits in a Dinur-Nissim database without noise
  addition
2
3 max_n_data = 1000;
4 min_n_data = 1000;
5 step_n_data = 10;
6
7 % number of random trials
8
9 n_trials = 100;
10
11 n_entry = floor((max_n_data-min_n_data)/step_n_data)+1;
12
13 n_q_recovery = zeros(1,n_entry);
14 n_d = zeros(1,n_entry);
15 n_q_norm = zeros(1,n_entry);
16
17 completion_counter_max = 10;% the consecutive number of times the min
  fraction correct is 1 before terminating the queryloop
18
19 i_noise = false; % set to false for no noise addition
20
21 i_entry = 0;
22
23 i_fig = 0;
24
25
26 for n_data = min_n_data:step_n_data:max_n_data
27
28     % noise level – we add gaussian noise with mean 0 and variance
      eta
29
30     sigma = sqrt(n_data)/2.0; % sigma for binomial distribution
31
32     eta = sigma*log(n_data); % ensuring the noise is just above the
      sqrt(n) growth
```

```

33
34
35 % query_fraction = linspace(1/n_data,1.0,query_max);
36
37 % generate random data set
38
39 d = randi([0,1],n_data,1);
40
41 options = optimset('display','off'); % turn off the display
42
43 % set the lower and upper bounds on the solution
44
45 lb = zeros(n_data,1);
46 ub = ones(n_data,1);
47
48 fraction_correct = zeros(n_trials,10000);
49
50 i_query = 0;
51
52 completion_counter = 0;
53
54 while (completion_counter < completion_counter_max)
55
56     i_query = i_query + 1;
57
58     max_fraction_corrrect = 0.0;
59     max_residual_norm = 0.0;
60
61     for i_trial = 1:n_trials
62
63         % generate the random query matrix
64
65         Q = randi([0,1], i_query, n_data);
66
67         % generate the query answers
68
69         ans_q = Q*d;
70
71         % add noise to the answers
72
73         rand_vec = normrnd(0,eta, [i_query, 1]);
74
75         if (i_noise)
76             ans_q = ans_q + rand_vec;
77         end

```

```

78
79     % now use constrained least squares to generate solution
80
81     [x_sol,res_norm,residual,exitflag,output] = lsqlin(Q,
82         ans_q,[],[],[],[],lb,ub, [], options);
83
84     max_residual_norm = max(max_residual_norm, res_norm);
85
86     % now round to 0 or 1
87
88     x_sol = round(x_sol);
89
90     % compute the percentage of bits returned correctly
91
92     n_correct = 0;
93
94     for i_bit = 1:n_data
95         if (abs(x_sol(i_bit) - d(i_bit)) <= 1.0e-3)
96             n_correct = n_correct + 1;
97         end
98     end
99
100     fraction_correct(i_trial, i_query) = n_correct/n_data;
101
102 end
103
104 max_fraction_correct = max(fraction_correct(:,i_query));
105 min_fraction_correct = min(fraction_correct(:,i_query));
106
107 if ((min_fraction_correct - 0.9) >= 0)
108     completion_counter = completion_counter + 1;
109 else
110     completion_counter = 0;
111 end
112
113 fprintf (' %5i trials n_data: %5i query: %5i comp_counter:
114         %5i min_fraction_correct %8.4e max_frac_correct %8.4e
115         max_residual: %8.4e \n', ...
116         n_trials, n_data, i_query, completion_counter,
117         min_fraction_correct, max_fraction_correct,
118         max_residual_norm)
119
120 end
121
122 n_query = i_query;

```

```

118
119 % now compute the mean percent correct and its variance
120
121 mean_fraction_correct = mean(fraction_correct);
122 var_fraction_correct = var(fraction_correct);
123
124 % now find the least value of query number that provides 100
    percent recovery
125
126 i_entry = i_entry+1;
127
128 n_d(i_entry) = n_data;
129
130 n_q_recovery(i_entry) = n_query;
131
132 for i = n_query:-1:1
133     if (abs(mean_fraction_correct(i) - 1) >= 1.0e-3)
134         n_q_recovery(i_entry) = i;
135         break;
136     end
137 end
138
139 % now produce a shaded distribution plot
140
141 x = 1:i_query;
142 y_mean = mean_fraction_correct(1:n_query);
143 y_10 = quantile(fraction_correct,0.10);
144 y_50 = quantile(fraction_correct,0.50);
145 y_90 = quantile(fraction_correct,0.90);
146
147 y_10 = y_10(1:n_query);
148 y_50 = y_50(1:n_query);
149 y_90 = y_90(1:n_query);
150
151
152 i_fig = i_fig+1;
153 figure(i_fig);
154 clf;
155
156 fprintf(' plotting figure %d...', i_fig);
157 hold on
158 plot(x,y_mean,'LineWidth',1.5);
159 plot(x,y_10);
160 plot(x,y_50);
161 plot(x,y_90);

```

```

162     hold off
163     title(['fraction correct vs. query for ', num2str(n_data), ' bits
           with ', num2str(n_trials), ' trials']);
164     drawnow;
165     fprintf (' plot complete\n')
166
167
168
169
170 end
171
172
173 % plot the min number of queries vs number of bits
174
175 i_fig = i_fig+1;
176
177 figure(i_fig);
178 clf;
179
180 plot (n_d(1:i_entry), n_q_recovery(1:i_entry));
181
182 drawnow;
183
184 % play with some possible normalizations of the min number of queries
185
186 for i_e = 1:i_entry
187     n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
188     %     n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
189 end
190
191 i_fig = i_fig+1;
192 figure(i_fig);
193 clf;
194
195 plot(n_d(1:i_entry), n_q_norm(1:i_entry));

```

Listing 2: Matlab script for Figures 5-2 and 5-3

```
1 % script to try to recover binary data set
2
3 max_n_data = 1000;
4 n_q_recovery = zeros(1,max_n_data);
5 n_d = zeros(1,max_n_data);
6 n_q_norm = zeros(1,max_n_data);
7
8 i_entry = 0;
9
10 for n_data = 100:100:max_n_data
11
12
13     max_query = n_data;
14     n_trials = 100;
15     query_percent = linspace(1/n_data,1.0,max_query);
16
17     % generate random data set
18
19     d = randi([0,1],n_data,1);
20
21     options = optimset('display','off'); % turn off the display
22
23     % set the lower and upper bounds on the solution
24
25     lb = zeros(n_data,1);
26     ub = ones(n_data,1);
27
28     percent_correct = zeros(n_trials,max_query);
29
30
31     for i_query = 1:1:max_query
32
33         fprintf (' n_data = %d   Performing query %d   with %d trials \
34                 \n', n_data, i_query, n_trials)
35
36         for i_trial = 1:n_trials
37
38             % generate the random query matrix
39
40             Q = randi([0,1], i_query, n_data);
41
42             % generate the query answers
43
```

```

44     ans_q = Q*d;
45
46     % now use constrained least squares to generate solution
47
48     [x_sol,res_norm,residual,exitflag,output] = lsqlin(Q,
49         ans_q,[],[],[],[],lb,ub, [], options);
50
51     % now round to 0 or 1
52
53     x_sol = round(x_sol);
54
55     % compute the percentage of bits returned correctly
56
57     n_correct = 0;
58
59     for i_bit = 1:n_data
60         if (abs(x_sol(i_bit) - d(i_bit)) <= 1.0e-3)
61             n_correct = n_correct +1;
62         end
63     end
64
65     percent_correct(i_trial, i_query) = n_correct/n_data;
66
67     end
68
69 end
70
71 % now compute the mean percent correct
72
73 min_percent_correct = min(percent_correct);
74 mean_percent_correct = mean(percent_correct);
75 var_percent_correct = 2.0*var(percent_correct); % note I'm taking
76     2 std devs
77 max_percent_correct = max(percent_correct);
78
79 % now find the lowest value of the number of queries that
80     provides 100 percent recovery
81
82 i_entry = i_entry+1;
83
84 n_d(i_entry) = n_data;
85
86 n_q_recovery(i_entry) = max_query;
87

```



```

86     for i = max_query:-1:1
87         if (abs(mean_percent_correct(i) - 1) >= 1.0e-3)
88             break;
89         else
90             n_q_recovery(i_entry) = n_q_recovery(i_entry) - 1;
91         end
92     end
93
94     % plot error bar plot
95
96     figure;
97
98     errorbar (mean_percent_correct, var_percent_correct)
99
100
101
102 end
103
104 % plot the min number of queries vs number of bits
105
106 figure;
107
108 plot (n_d(1:i_entry), n_q_recovery(1:i_entry));
109
110 % play with some possible normalizations of the min number of queries
111     -
112 % here we try direct proportionality to number of bits
113
114 for i_e = 1:i_entry
115     n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
116 end
117
118 figure;
119
120 plot(n_d(1:i_entry), n_q_norm(1:i_entry));

```

Listing 3: Matlab script for Figure 5-4

```
1 % script to examine the distribution of number of bits recovered for
  a
2 % fixed number of random bits in a database
3
4 max_n_data = 10;
5 min_n_data = 100;
6 step_n_data = 10;
7
8 % number of random trials
9
10 n_trials = 100;
11
12 n_entry = floor((max_n_data-min_n_data)/step_n_data)+1;
13
14 n_q_recovery = zeros(1,n_entry);
15 n_d = zeros(1,n_entry);
16 n_q_norm = zeros(1,n_entry);
17
18 completion_counter_max = 10;% the consecutive number of times the min
  fraction correct is 1 before terminating the queryloop
19
20 i_noise = true; % set to false for no noise addition
21
22 i_entry = 0;
23
24 i_fig = 0;
25
26
27 for n_data = min_n_data:step_n_data:max_n_data
28
29     % noise level – we add gaussian noise with mean 0 and variance
      eta
30
31     sigma = sqrt(n_data)/2.0; % sigma for binomial distribution
32
33     eta = sigma*log(n_data); % ensuring the noise is just above the
      sqrt(n) growth
34
35
36     % generate random data set
37
38     d = randi([0,1],n_data,1);
39
40     options = optimset('display','off'); % turn off the display
```

```

41
42 % set the lower and upper bounds on the solution
43
44 lb = zeros(n_data,1);
45 ub = ones(n_data,1);
46
47 fraction_correct = zeros(n_trials,10000);
48
49 i_query = 0;
50
51 completion_counter = 0;
52
53 while (completion_counter < completion_counter_max)
54
55     i_query = i_query + 1;
56
57     max_fraction_corrrect = 0.0;
58     max_residual_norm = 0.0;
59
60     for i_trial = 1:n_trials
61
62         % generate the random query matrix
63
64         Q = randi([0,1], i_query, n_data);
65
66         % generate the query answers
67
68         ans_q = Q*d;
69
70         % add noise to the answers
71
72         rand_vec = normrnd(0,eta, [i_query, 1]);
73
74         if (i_noise)
75             ans_q = ans_q + rand_vec;
76         end
77
78         % now use constrained least squares to generate solution
79
80         [x_sol,res_norm,residual,exitflag,output] = lsqin(Q,
81             ans_q,[],[],[],[],lb,ub, [], options);
82
83         max_residual_norm = max(max_residual_norm, res_norm);
84
85         % now round to 0 or 1

```

```

85
86     x_sol = round(x_sol);
87
88     % compute the percentage of bits returned correctly
89
90     n_correct = 0;
91
92     for i_bit = 1:n_data
93         if (abs(x_sol(i_bit) - d(i_bit)) <= 1.0e-3)
94             n_correct = n_correct +1;
95         end
96     end
97
98     fraction_correct(i_trial, i_query) = n_correct/n_data;
99
100    end
101
102    max_fraction_correct = max(fraction_correct(:,i_query));
103    min_fraction_correct = min(fraction_correct(:,i_query));
104
105    if ((min_fraction_correct - 0.9) >= 0)
106        completion_counter = completion_counter + 1;
107    else
108        completion_counter = 0;
109    end
110
111    fprintf (' %5i trials n_data: %5i query: %5i comp_counter:
112            %5i min_fraction_correct %8.4e max_frac_correct %8.4e
113            max_residual: %8.4e \n', ...
114            n_trials, n_data, i_query, completion_counter,
115            min_fraction_correct, max_fraction_correct,
116            max_residual_norm)
117
118    end
119
120    n_query = i_query;
121
122    % now compute the mean percent correct and its variance
123
124    mean_fraction_correct = mean(fraction_correct);
125    var_fraction_correct = var(fraction_correct);
126
127    % now find the least value of query number that provides 100
128    percent recovery

```

```

125     i_entry = i_entry+1;
126
127     n_d(i_entry) = n_data;
128
129     n_q_recovery(i_entry) = n_query;
130
131     for i = n_query:-1:1
132         if (abs(mean_fraction_correct(i) - 1) >= 1.0e-3)
133             n_q_recovery(i_entry) = i;
134             break;
135         end
136     end
137
138     % now produce a shaded distribution plot
139
140     x = 1:i_query;
141     y_mean = mean_fraction_correct(1:n_query);
142     y_10 = quantile(fraction_correct,0.10);
143     y_50 = quantile(fraction_correct,0.50);
144     y_90 = quantile(fraction_correct,0.90);
145
146     y_10 = y_10(1:n_query);
147     y_50 = y_50(1:n_query);
148     y_90 = y_90(1:n_query);
149
150
151     i_fig = i_fig+1;
152     figure(i_fig);
153     clf;
154
155     fprintf(' plotting figure %d...', i_fig);
156     hold on
157     plot(x,y_mean,'LineWidth',1.5);
158     plot(x,y_10);
159     plot(x,y_50);
160     plot(x,y_90);
161     hold off
162     title(['fraction correct vs. query for ', num2str(n_data),' bits
163           with ',num2str(n_trials),' trials']);
164     drawnow;
165     fprintf (' plot complete\n')
166
167
168

```

```
169 end
170
171
172 % plot the min number of queries vs number of bits
173
174 i_fig = i_fig+1;
175
176 figure(i_fig);
177 clf;
178
179 plot (n_d(1:i_entry), n_q_recovery(1:i_entry));
180
181 drawnow;
182
183 % play with some possible normalizations of the min number of queries
184
185 for i_e = 1:i_entry
186     n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
187     %     n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
188 end
189
190 i_fig = i_fig+1;
191 figure(i_fig);
192 clf;
193
194 plot(n_d(1:i_entry), n_q_norm(1:i_entry));
```

Listing 4: Matlab script for Figure 6-6

```
1 % script to examine the accuracy of a sum query as a function of the
  value
2 % of epsilon
3
4 n_data_row = [100 200 500 1000 2000 5000];
5
6 % number of random trials
7
8 n_trials = 1000;
9
10 trial_result = zeros(n_trials,1);
11
12 % the set of privacy loss parameters we wish to examine
13
14 eps_row = [0.001 0.005 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09
  0.1 0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.2 0.3 0.4 0.5
  0.6 0.7 0.8 0.9 1.0 ];
15
16
17 n_d_entry = length(n_data_row);
18 n_e_entry = length(eps_row);
19
20
21 query_accuracy = zeros(n_d_entry, n_e_entry); %
22
23
24 for i_d_entry = 1:n_d_entry % loop over the values of the number of
  bits
25
26     n_data = n_data_row(i_d_entry);
27
28     fprintf (' number of data bits: %d \n ', n_data);
29
30     for i_e_entry = 1:n_e_entry % loop over the values of epsilon
31
32         epsilon = eps_row(i_e_entry);
33
34         % noise level – we add gaussian noise with mean 0 and
          variance eta
35
36         eta = 2/epsilon^2; % this sets the variance to the
          equivalent of the two sided exponential
37
38         for i_trial = 1:n_trials % do a number of trials to get
```

```

39         reasonable statistics
40         % generate random data set
41
42         d = randi([0,1],n_data,1);
43
44         % compute the correct sum
45
46         sum_query = sum(d);
47
48         % add noise to the sum of the data set – here we add a
           Laplace
49         % distribution with parameter epsilon
50
51         unif = rand() - 0.5;
52         laplace_rand_var = -1./epsilon*sign(unif)*log(1-2*abs(
           unif));
53
54     %         rand_num = normrnd(0,sqrt(eta), [1, 1]);Q_n
55
56         rand_num = laplace_rand_var;
57         noised_sum = round(sum_query + rand_num);
58
59         trial_result(i_trial) = 1.0 - abs((noised_sum-sum_query)
           /sum_query); % accuray – 1 is perfect and then it
           decreases as error decreases
60
61     end
62     mean_error = mean(trial_result);
63
64     fprintf ('         epsilon = %d variance = %d mean_error=%d\n',
           epsilon, eta, mean_error);
65
66     query_accuracy(i_d_entry,i_e_entry) = mean_error;
67
68     end
69
70 end
71
72
73
74 % now plot the results
75
76 figure;
77

```



```

78 hold on
79
80 for i_curve = 1:n_d_entry
81
82     x = eps_row;
83
84     y = query_accuracy(i_curve, 1:n_e_entry);
85
86     plot (x,y);
87
88 end
89
90 % set the axes – anything below a query accuracy of 0.0 is pretty
    useless
91 axis([0 1.0 0 1.01]);
92
93 % form the legend
94
95 for i_curve = 1:n_d_entry
96     legendCell{i_curve} = num2str(n_data_row(i_curve), 'N =%-d');
97 end
98
99 legend(legendCell);
100
101 % label the axes
102
103 xlabel('Privacy loss parameter – \epsilon');
104 ylabel('Query accuracy');
105
106 % title the plot
107
108 title(' Dinur–Nissim query accuracy vs privacy loss parameter \
    epsilon');

```

Listing 5: Matlab script for Figure 6-7

```
1
2 % Matlab script to examine what percentage of bits are recovered for
  a given
3 % privacy loss parameter and a given number of queries in the
  presence of
4 % noise. We use a two-sided Laplace distribution to sample the noise.
5
6 % the set of database size we wish to examine
7
8 n_data_row = [4000];
9
10 % number of random trials
11
12 n_trials = 10;
13
14 trial_fraction_correct = zeros(n_trials,1);
15
16 % the set of privacy loss parameters we wish to examine
17
18 eps_row = [ 0.01 0.02 0.03 0.04 0.05 0.1 0.2 0.25 0.3 0.4 0.5 1.0  ];
19
20 % the set of multiples of the number of data points we have that we
  wish to examine
21
22 n_mult_row = [1 5 10 20];
23
24 n_d_entry = length(n_data_row);
25 n_e_entry = length(eps_row);
26 n_m_entry = length(n_mult_row);
27
28 options = optimset('display','off'); % turn off the display for the
  optimizer
29
30 % array of fraction of number of bits correct as a function of number
  of bits, number of queries, and epsilon
31 fraction_correct = zeros(n_d_entry, n_m_entry, n_e_entry);
32
33 % loop over the values of the number of bits
34 for i_d_entry = 1:n_d_entry
35
36     n_data = n_data_row(i_d_entry);
37
38     fprintf (' number of data bits: %d \n ', n_data);
39
```

```

40 % set the lower and upper bounds on the solution
41
42 lb = zeros(n_data,1);
43 ub = ones(n_data,1);
44
45 % generate random data set
46
47 d = randi([0,1],n_data,1);
48
49 % loop over the values of epsilon
50 for i_e_entry = 1:n_e_entry
51
52     epsilon = eps_row(i_e_entry);
53
54     % noise level – we add Laplace noise with mean 0 and variance
55     % this sets the variance to the equivalent of the two sided
56     % exponential
57     eta = 2/epsilon^2;
58
59     fprintf ('      epsilon = %d variance = %d \n', epsilon, eta)
60     ;
61
62     % loop over the queries – we do various multiples of the
63     % number of
64     % data points
65
66     for i_m_entry = 1:n_m_entry
67
68         i_query = n_data*n_mult_row(i_m_entry);
69
70         % we do n_trials trials and average the results
71
72         max_residual_norm = 0;
73
74         for i_trial = 1:n_trials
75
76             % generate the random query matrix
77
78             Q = randi([0,1], i_query, n_data);
79
80             % generate the query answers
81
82             ans_q = Q*d;

```

```

81         % add noise to the answers
82
83         % add noise to the sum of the data set – here we add
           a Laplace
84         % distribution with parameter epsilon
85
86         unif = rand(i_query,1) - 0.5;
87         laplace_rand_var = -1./epsilon.*sign(unif).*log(1-2*
           abs(unif));
88
89         ans_q = ans_q + laplace_rand_var;
90
91         % now use constrained least squares to generate
           solution
92
93         [x_sol,res_norm,residual,exitflag,output] = ...
           lsqlin(Q,ans_q,[],[],[],[],lb,ub, [], options);
94
95         max_residual_norm = max(max_residual_norm, res_norm);
96
97         % now round to 0 or 1
98
99         x_sol = round(x_sol);
100
101         % compute the percentage of bits returned correctly
102
103         n_correct = 0;
104
105         for i_bit = 1:n_data
106             if (abs(x_sol(i_bit) - d(i_bit)) <= 1.0e-3)
107                 n_correct = n_correct +1;
108             end
109         end
110
111         trial_fraction_correct(i_trial) = n_correct/n_data;
112
113     end
114
115     max_fraction_correct = max(trial_fraction_correct);
116     min_fraction_correct = min(trial_fraction_correct);
117     mean_fraction_correct = mean(trial_fraction_correct);
118     var_fraction_correct = var(trial_fraction_correct);
119
120     fprintf ('           n_data: %5i query: %5i
121             mean_fraction_correct %8.4e   max_residual: %8.4e \n',

```

```

122         ...
           n_data, i_query, mean_fraction_correct,
           max_residual_norm)
123
124         fraction_correct(i_d_entry,i_m_entry,i_e_entry) =
           mean_fraction_correct;
125
126     end
127 end
128 end
129
130 % now plot the results
131
132 [X, Y] = meshgrid(n_mult_row, eps_row);
133
134 % loop over the size of the data vector
135
136 Z = zeros(n_e_entry, n_m_entry);
137
138 for i_d_entry = 1:n_d_entry
139
140     for i_e_entry = 1:n_e_entry
141
142         for i_m_entry = 1:n_m_entry
143
144             Z(i_e_entry, i_m_entry) = fraction_correct(i_d_entry,
               i_m_entry, i_e_entry); % load the array of results for
               each data set size
145
146         end
147
148     end
149
150     figure;
151     surf(X,Y,Z);
152     set(gca,'XScale','linear')
153     set(gca,'YScale','linear')
154 end

```

---

## References

- [1] John M Abowd. Staring Down the Database Reconstruction Theorem. Presentation to AAAS Annual Meeting Feb 16, 2019, 2019.
- [2] Robert Ashmead. Estimating the Variance of Complex Differentially Private Algorithms. Presentation to Joint Statistical Meetings, American Statistical Association, July 27, 2019, 2019.
- [3] Raj Chetty and John Friedman. A practical method to reduce privacy loss when disclosing statistics based on small sample. *American Economic Review Papers and Proceedings*, 109:414–420.
- [4] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. Constrained Differential Privacy for Count Data. *arXiv e-prints*, page 1710.00608, Oct 2017.
- [5] Irit Dinur and Kobbi Nissim. Revealing Information while Preserving Privacy. In *PODS*, pages 202–210. ACM, 2003.
- [6] Cynthia Dwork and Jing Lei. Differential Privacy and Robust Statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, pages 371–380. Association for Computing Machinery, 2009.
- [7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.
- [8] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Found. Trends. Theor. Comput. Sci.*, 9(3-4):211–407, 2013.
- [9] Ivan P. Fellegi and Alan B. Sunter. A Theory for Record Linkage. *J. Am. Stat. Assoc.*, 64(328):1183–1210, 1969.
- [10] Simson Garfinkel, John M. Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data. *Commun. ACM*, 62(3):46–53, 2019.

- 
- [11] Solomon W. Golomb and Guang Gong. *Signal design for good correlation*. Cambridge, 2005.
- [12] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2019.
- [13] Mark Hansen. To Reduce Privacy Risk, Census Plans to Report Less Accurate Data. *New York Times*, Dec. 6, 2018.
- [14] D. Kifer. Design Principles of the TopDown Algorithm. Presentation to JASON.
- [15] Ios Kotsogiannis, Yuchao Tao, Ashwin Machanavajjhala, Gerome Miklau Umass, and Amherst Michael Hay. Architecting a Differentially Private SQL Engine. <http://cidrdb.org/cidr2019/papers/p125-kotsogiannis-cidr19.pdf>.
- [16] Philip Leclerc. Generating Microdata with Complex Invariants under Differential Privacy. Presentation to Joint Statistical Meeting, American Statistical Association, 2019.
- [17] Philip Leclerc. Results from a Consolidated Database Reconstruction and Intruder Re-identification Attack on the 2010 Decennial Census. Presentation at Workshop "Challenges and New Approaches for Protecting Privacy in Federal Statistical Programs", 2019.
- [18] Justin Levitt. Uses of 2020 Redistricting Data. Presentation to JASON.
- [19] Chao Li, Michael Hay, Gerome Miklau, and Yue Wang. A Data- and Workload-Aware Algorithm for Range Queries Under Differential Privacy. <http://arxiv.org/abs/1410.0265>, 2014.
- [20] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *VLDB J.*, 24(6):757–781, 2015.
- [21] Ashwin Machanavajjhala. Interpreting Differential Privacy. Presentation to JASON.
- [22] David J. C. MacKay. *Information Theory, Inference, & Learning*. Cambridge University Press, 2003.

- 
- [23] Rachel Marks. How the 2020 Census Products Reflect Data user Feedback. Presentation to JASON.
- [24] Laura McKenna. Disclosure Avoidance for the 1970-2010 Censuses, 2018. <https://www2.census.gov/ces/wp/2018/CES-WP-18-47.pdf>.
- [25] Ryan McKenna, Gerome Miklau, Michael Hay, and Ashwin Machanavajjhala. Optimizing Error of High-dimensional Statistical Queries Under Differential Privacy. *Proc. VLDB Endow.*, 11(10):1206–1219, June 2018.
- [26] Kobbi Nissim, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, David R O'brien, and Salil Vadhan. Differential privacy: a primer for a non-technical audience. *Vanderbilt J. Entertain. Technol. Law*, page 1021596, 2018.
- [27] A. Ramachandran, L. Singh, E. Porter, and F. Nagle. Exploring Re-Identification Risks in Public Domains. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*, 2012.
- [28] Jerome P. Reiter. Differential Privacy and Federal Data Releases. *Annu. Rev. Stat. Its Appl.*, 6(1):85–101, 2018.
- [29] Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. Differential Privacy and Census Data: Implications for Social and Economic Research. *AEA Pap. Proc.*, 109:403–408, 2019.
- [30] William Sexton. Disclosure Avoidance At-Scale. Presentation to JASON.
- [31] C. E. Shannon. Communication in the presence of noise. *Proc. Inst. Radio Engineers*, 37(1):10–21, 1949.
- [32] Latanya Sweeney, Merce Crosas, and Michael Bar-Sinai. Sharing Sensitive Data with Confidence: the Datatags System. *Technol. Sci.*, pages 1–34, 2015.
- [33] US Census Bureau. Census Bureau Continues to Boost Data Safeguards. <https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html>.
- [34] US Census Bureau. Census End to End Disclosure Avoidance System. <https://github.com/uscensusbureau/census2020-das-e2e>, 2019.



- 
- [35] US Census Bureau. Census Population Density by County. <https://www.census.gov/library/visualizations/2010/geo/population-density-county-2010.html>, 2019.
- [36] D. van Riper. Differential Privacy and the Decennial Census. Presentation to JASON.
- [37] David van Riper. Differential Privacy and the Decennial Census. [https://assets.ipums.org/\\_files/intro\\_to\\_differential\\_privacy\\_IPUMS\\_workshop.pdf](https://assets.ipums.org/_files/intro_to_differential_privacy_IPUMS_workshop.pdf), 2019.
- [38] Victoria Velkoff. Proposed 2020 Census Data Products. Presentation to JASON.
- [39] James Whitehorne. Overview of redistricting data products. Presentation to JASON.
- [40] Tommy Wright. Suitability Assessment of Data treated by DA Methods for Redistricting: Observations. Presentation to JASON.



**Comment on “The Impact of the U.S. Census Disclosure Avoidance System on Redistricting and Voting Rights Analysis,” by [Kenny et al.](#)**

**Sam Wang and Ari Goldbloom-Helzner  
Electoral Innovation Lab, Green Hall, Princeton University, Princeton, NJ 08544.**

**June 2, 2021**

The plaintiffs in *Alabama v. Department of Commerce* (case no. 3:21-cv-00211-RAH-ECM-KCN) have filed a supplemental statement to their expert report in which they attach a working paper by Christopher Kenny and other students, working in collaboration with Professor Kosuke Imai of Harvard University. Prof. Imai is a recognized expert in automated methods for drawing district maps. Kenny et al. report results of applying ensemble simulation methods to the Census Bureau’s DAS 12.2-demonstration data set, in which noise was added to 2010 Census data. For comparison they do calculations using the Census 2010 data release, in which privacy protection was accomplished using swapping, an older method of disclosure avoidance. Kenny et al. report differences between their simulations under the two conditions, and conclude that these differences arise from bias. They assert that these biases are large enough to make it difficult to comply with redistricting requirements. They conclude that such issues can be avoided by reverting to the swapping method or by suppressing some block-level Census tables.

This working paper has not been through peer review. We therefore performed our own examination of the manuscript. Our group at Princeton University, the Electoral Innovation Lab, is expert in analysis of election and redistricting data. One of our projects, the Princeton Gerrymandering Project, does ensemble analysis in its own work, and we are published in this area. We are therefore qualified to comment on the work of Kenny et al.

In our reading, we encountered four major problems that cast doubt on the conclusions.

**1. The algorithm is unreviewed and adds unnecessary complexity to the analysis**

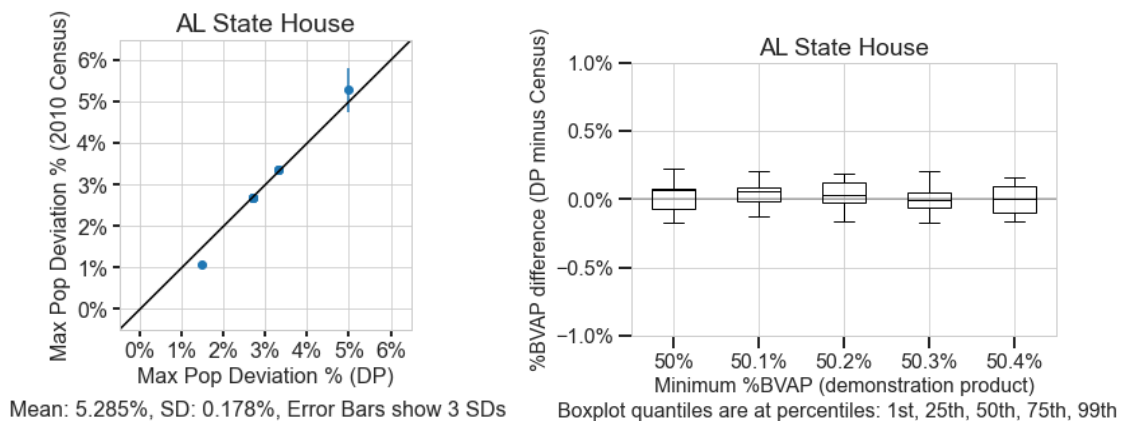
The practical question relating to the DAS 12.2-data set is whether its use would affect the properties of real districts: school districts, county commissioner districts, legislative districts, Congressional districts, and so on. This work does not calculate that. Instead, it samples the properties of ensembles of many simulated districts, a process that explores the entire range of possibilities, including outliers.

Kenny et al. are using a new sampling algorithm, which has not been through peer review. There is an unknown risk that some of the results arise from unknown characteristics of the sampling algorithm. These results would be of interest to researchers but not necessarily redistricters. For example, the effects in Figure 1 and 2 appear to depend on the max() operation, which is sensitive to the properties of outliers of the algorithm. Likewise, the effects shown in Sections 4 through 7 might also depend on peculiarities of the algorithm.

**2. When ensemble analysis is done properly, differences from Census 2010 data are of no practical consequence**

There is also some question as to whether the algorithm has been applied properly to the question at hand. Some of the reported findings may arise as a chance result of running the same algorithm twice, which at one point (Section 6) is how they compare the effects of DAS 12.2-data and 2010 Census data. A better approach would be to run one simulation and test the different datasets under the same set of maps.

We have performed such a simulation using a widely-accepted redistricting algorithm, GerryChain. One such example of our work is shown below. In this work, we simulated 10,000 Alabama state House districts. We then calculated two key parameters of redistricting: maximum population deviation, which is allowed to be up to +/-5% of the average for non-Congressional legislative districts;<sup>1</sup> and the percentage Black voting-age population (“%BVAP”). We calculated both of these quantities using the DAS 12.2-data, and compared them with the same quantities for the exact same districts with the 2010 Census data. In this way, we were able to calculate the difference that was made by using demonstration DAS 12.2-data and the Census 2010 data release for 10,000 specific state district plans. The results are shown in slide #8 of our presentation at <http://bit.ly/SamWang-Princeton-Census-privacy> and are reproduced here:



<sup>1</sup> *Brown v. Thomson*, 462 U.S. 835 (1983)

We found that a plan's maximum population deviation was very similar under both conditions. The left-hand graph shows that the maximum population deviation was within a fraction of a percentage point when comparing DAS 12.2-data with Census 2010 data. In short, the two datasets perform nearly identically for purposes of one person, one vote analysis.

It should also be noted that Kenny et al. have erroneously stated the one person, one vote principle in the case of Congressional districts as mandating exact population equality to within one person. *Tennant vs. Jefferson County Commission*, 567 U.S. 758, (2012) found that a population variance of 0.79% was acceptable in light of a legitimate redistricting objective.

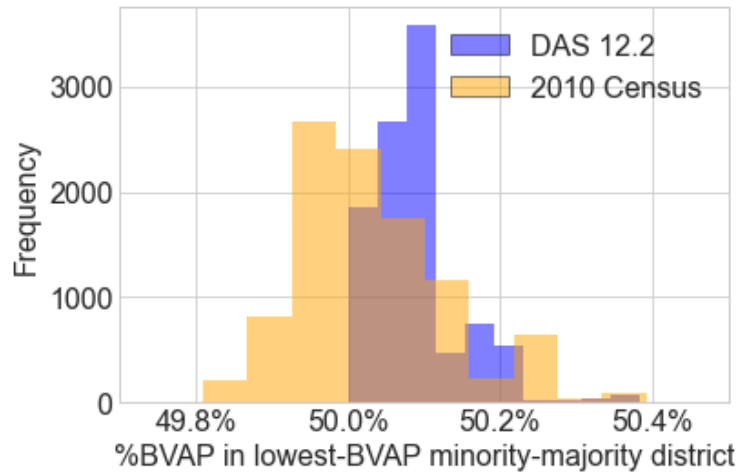
In regard to %BVAP, we find that for districts that were in the range of 50.0-50.5% BVAP, calculations based on DAS 12.2-data and Census 2010 data were closer than 0.1 percentage point (as a fraction of total voting-age population) in the majority of cases, and always closer than 0.2 percentage point. Such small differences are of no practical consequence for assessing the voting performance of a district.

Kenny et al. have taken a different approach to calculating %BVAP which we contend is misleading. They used a hard threshold of 50% when defining majority-minority districts. Their statistics are predicted on the idea that if a district were 50.1% BVAP with the DAS 12.2-data, but only 49.9% BVAP with the 2010 Census swapped data, the voting performance of the district would be meaningfully different. In fact, no performance difference would result. Furthermore, their analysis fails to recognize that a nearly-equal number of districts would change by a similar amount in the opposite direction. Any evaluation on hard thresholds is bound to lead to some maps that would have been over 50%, and instead are just under, or vice versa. This difference is functionally meaningless.<sup>2</sup>

The same error appears in Section 7, Table 2, where Kenny et al. examine school districts and evaluate on different datasets. They find discrepancies on the number of majority-minority districts. However, the use of a hard threshold of 50% makes it impossible to know whether this result is exaggerated. As a demonstration, our own simulations generated this distribution of %BVAP in Alabama state House districts:

---

<sup>2</sup> The Voting Rights Act also protects minority representation by allowing the use of opportunity-to-elect districts, in which minority groups are large enough to play a dominant role in the primary of a party that is likely to win the general election. Research in political science shows that such a district arises if the percentage of minority voting age population falls between 30 and 50 percent.



Since there were 27 districts in the Alabama state House that had %BVAP greater than 50% with the DAS 12.2-data, we studied the 27th most %BVAP district which is the lowest-%BVAP district that still qualifies as a majority-minority district. In our 10,000 map ensemble, we find that, at worst, the 27th most %BVAP district under the 2010 Census data would have had 49.8% BVAP. This difference is minute and would not have practical effects on representation.

**3. The calculated effects on partisan performance are of no consequence for real districting situations**

Many of the ensemble effects reported are quite small. Figure 4 shows the distribution of performance of thousands of plans. However, the average performance of the distributions is not shown clearly. To the extent that a difference can be seen in this figure, the difference appears to be well under half of one Congressional district. That difference refers to the average of thousands of plans.

However, real-life districting consists of drawing a single plan. This process is under the control of human beings in every state and jurisdiction in the United States. It is well-known that the human-led redistricting process can have effects that exceed one seat in magnitude. Therefore the average properties of an ensemble are inconsequential for evaluating the properties of Census data.

**4. Comparisons with Census 2010 data make a fundamentally flawed assumption about ground truth**

Finally, Kenny et al. make a conceptual error of a type that runs through many arguments made in this case. That error consists of the assumption that Census 2010 data is ground truth. This assumption is categorically false. The Census 2010 data release itself used an older method of disclosure avoidance, swapping, to move racial characteristics around. Sections 6 and 7 are marred by this assumption.

This is not a trivial error. The 2010 Census data is itself inaccurate to some extent; it is not ground truth. Data swapping is [known to alter](#) the apparent minority population in areas where that population is scarce. Thus, it is impossible to determine the actual racial characteristics of districts drawn using 2010 Census data. Ironically, the new proposed method is more rigorous and lends itself more easily to quality control.

Kenny et al. recommend for the Bureau to rely on the swapping method for its Disclosure Avoidance System instead of differential privacy; however, they do not evaluate the effects of swapping on privacy or representation. In fact, the Bureau's [recent study on swapping](#) found it to be unequivocally worse than the current DAS in its impact on re-identification and accuracy metrics.

It should also be noted that the Census count itself is prone to inaccuracies<sup>3</sup>that disproportionately affect minority communities. The 2010 Census undercounted 2.1 percent of the Black population and 1.5 percent of the Hispanic population. Nonetheless, the 2010 Census was accepted by the courts and used as the basis for redistricting litigation. Indeed, the Alabama plaintiffs find it clear that “past methods [swapping] do not violate the Secretary's obligations to report accurate ‘tabulations of population under § 141(c).’”<sup>4</sup> In short, known errors of tabulation far exceed any consequences of disclosure avoidance. In other words, if the 2010 Census data was considered fit for use, the DAS 12.2 approach performs equivalently to below the limits of detectability.

>>>

We would like to close with a statement of a general principle which can guide the court in understanding the arcane-seeming subject of adding noise to Census data. This point, a general one regarding counting statistics, provides a general framework for thinking about the use of Disclosure Avoidance System protections.

The addition of random error to Census data is fundamentally different from systematic error. An example of systematic error is undercounting that occurs everywhere. Such error is indeed detrimental to the accurate counting of persons: a 5% undercount in each block leads to a 5% undercount in the entire population.

Random error is fundamentally different. Random errors tend to cancel one another out. As a general rule of thumb, that cancellation has “square-root” properties. For example, combining 100 blocks would tend to make percentage errors square-root-of-100, or 10 times, smaller. This fundamental principle allows measures of individual blocks to be uncertain, while allowing measures of aggregates such as districts to be highly accurate.

---

<sup>3</sup> [https://www.census.gov/newsroom/releases/archives/2010\\_census/cb12-95.html](https://www.census.gov/newsroom/releases/archives/2010_census/cb12-95.html)

<sup>4</sup> Plaintiff's Motion for a Preliminary Injunction, p. 42, Mar. 11, 2021

In light of this general principle, it is worthwhile to look at other recent work. Nothing in the Kenny et al. working paper addresses our recent findings that estimates on very small populations may indeed be affected by DAS 12.2-data, but larger ones are not affected. Nor do Kenny et al. address a recent article by [Cohen, Duchin et al.](#), which finds that racial polarization analysis is unaffected by the addition of privacy-protecting error.

## **Conclusion**

The dramatic claims of Kenny et al. about functional consequences of disclosure avoidance should be regarded with skepticism, at least until the work has passed peer review.

Samuel Wang, Ph.D.  
Electoral Innovation Lab  
Professor, Princeton University

Ari Goldbloom-Helzner  
Electoral Innovation Lab  
Data Analyst, Princeton University

-- Last Updated 6/2/2021 3:20PM

# Understanding the 2020 Census Disclosure Avoidance System:

## *Differential Privacy 101*

### **Michael Hawes**

Senior Advisor for Data Access and Privacy  
Research and Methodology Directorate  
U.S. Census Bureau

May 4, 2021

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
**2020**



Webinar Series:

## Understanding the 2020 Census Disclosure Avoidance System

All webinars start at **1:00 pm EDT**

No pre-registration necessary. We will archive recordings to the website.

\*Search “*Disclosure Updates*” at [www.census.gov](http://www.census.gov)

Or link: <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html>

Day	Date	Title
T	May 4	Differential Privacy 101
F	May 7	The Census Bureau's Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census
Th	May 13	Differential Privacy 201 and the TopDown Algorithm
F	May 14	Highlights of the April 2021 Detailed Summary Metrics
F	May 21	Analysis of April 2021 Demonstration Data for Redistricting and Voting Rights Act Use Cases

2020CENSUS.GOV

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Acknowledgements

**This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, including:** John Abowd, Tammy Adams, Robert Ashmead, Craig Corl, Ryan Cummings, Jason Devine, John Fattaleh, Simson Garfinkel, Nathan Goldschlag, Michael Hawes, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Kyle Irimata, Dan Kifer, Philip Leclerc, Ashwin Machanavajhala, Christian Martindale, Gerome Miklau, Claudia Molinar, Brett Moran, Ned Porter, Sarah Powazek, Vikram Rao, Chris Rivers, Anne Ross, Ian Schmutte, William Sexton, Rob Sienkiewicz, Matthew Spence, Tori Velkoff, Lars Vilhuber, Bei Wang, Tommy Wright, Bill Yates, and Pavel Zhurlev.

**For more information and technical details relating to the issues discussed in these slides, please contact the author at [michael.b.hawes@census.gov](mailto:michael.b.hawes@census.gov).**

**Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.**

**The statistics included in this newsletter have been cleared for public dissemination by the Census Bureau's Disclosure Review Board (CBDRB-FY20-DSEP-001, CBDRB-FY20-281, and CBDRB-FY20-101).**

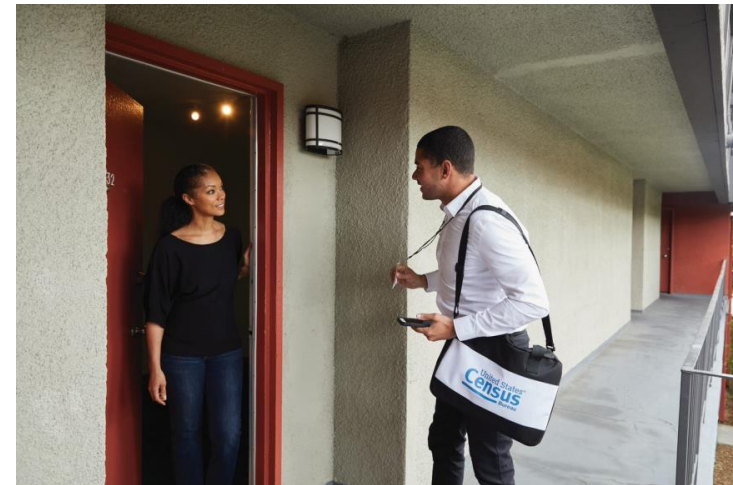
Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Our Commitment to Privacy and Confidentiality

Data stewardship is central to the Census Bureau's mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



# Upholding our Promise: Today and Tomorrow

**We cannot merely consider privacy threats that exist today.**

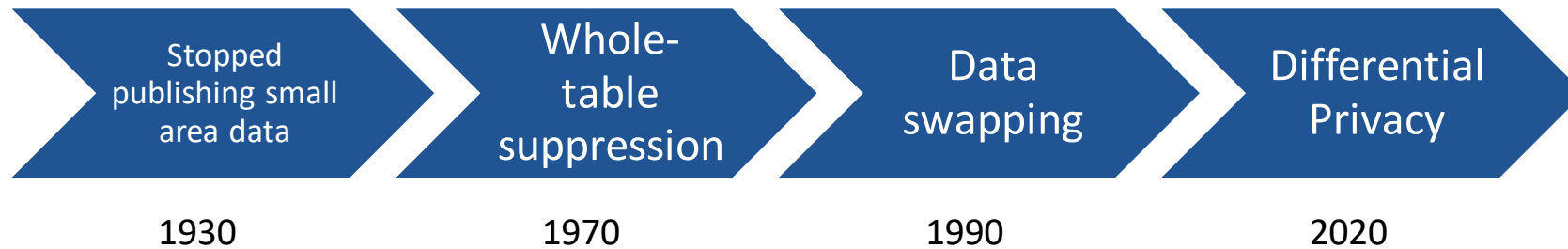
**We must ensure that our disclosure avoidance methods are also sufficient to protect against the threats of tomorrow!**



# The Census Bureau's Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.



# The Privacy Challenge

Every time you release any statistic calculated from a confidential data source you “leak” a small amount of private information.

If you release too many statistics, too accurately, you will eventually reveal the entire underlying confidential data source.

*Dinur, Irit and Kobbi Nissim (2003) “Revealing Information while Preserving Privacy” PODS, June 9-12, 2003, San Diego, CA*

7



Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# The Growing Privacy Threat

## More Data and Faster Computers!

In today's digital age, there has been a proliferation of databases that could potentially be used to attempt to undermine the privacy protections of our statistical data products.

Similarly, today's computers are able to perform complex, large-scale calculations with increasing ease.

These parallel trends represent new threats to our ability to safeguard respondents' data.

# Reconstruction

The recreation of individual-level data from tabular or aggregate data.

If you release enough tables or statistics, eventually there will be a unique solution for what the underlying individual-level data were.

Computer algorithms can do this very easily.

	4					2	
			7				4
1		7	8			5	
			9			3	8
5							
			6		8		
3						4	5
	8	5				1	9
		9		7	1		



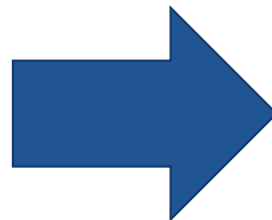
# Reconstruction: An Example



	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

# Reconstruction: An Example

	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7



Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married
36	Female	Black	Married
8	Female	Black	Single
18	Male	White	Single
24	Female	White	Single

This table can be expressed by 164 equations.  
Solving those equations takes 0.2 seconds on a 2013  
MacBook Pro.

# Re-identification

Linking public data to external data sources to re-identify specific individuals within the data.

Name	Age	Sex		Age	Sex	Race	Relationship
Jane Smith	66	Female	+	66	Female	Black	Married
Joe Public	84	Male		84	Male	Black	Married
John Citizen	30	Male		30	Male	White	Married

External Data

Confidential Data

# Reconstructing the 2010 Census

- The 2010 Census collected information on the age, sex, race, ethnicity, and relationship (to householder) status for ~309 Million individuals. (1.9 Billion confidential data points)
- The 2010 Census data products released over 150 billion statistics
- We conducted an internal experiment to see if we could reconstruct and re-identify the 2010 Census records.



# Reconstructing the 2010 Census: What Did We Find?

1. On the 309 million reconstructed records, census block and voting age (18+) were correctly reconstructed for all individuals in all 6,207,027 inhabited blocks.
2. Block, sex, age (in years), race (OMB 63 categories), and ethnicity were reconstructed:
  1. Exactly for 46% of the population (142 million individuals)
  2. Within +/- one year for 71% of the population (219 million individuals)
3. Block, sex, and age were then linked to commercial data, which provided presumed re-identification of 45% of the population (138 million individuals).
4. Name, block, sex, age, race, ethnicity were then compared to the confidential data, which yielded confirmed re-identifications for 38% of the presumed re-identifications (52 million individuals).
5. For the confirmed re-identifications, race and ethnicity are learned correctly, though the attacker may still have uncertainty.

# The Census Bureau's Decision

- Advances in computing power and the availability of external data sources make database reconstruction and re-identification increasingly likely.
- The Census Bureau recognized that its traditional disclosure avoidance methods are increasingly insufficient to counter these risks.
- To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections.



# Disclosure Avoidance

Disclosure avoidance methods seek to make reconstruction and re-identification more difficult, by:

- Reducing precision
- Removing vulnerable records, or
- Adding uncertainty

Commonly used (legacy) methods include:

- Complementary suppression
- Rounding
- Top/Bottom coding of extreme values
- Sampling
- Record swapping
- Noise injection

# Problem #1 – Impact on Data

All statistical techniques to protect privacy impose a tradeoff between the **degree of privacy protection** and the resulting **accuracy of the data**.

Swap rates, noise injection parameters, cell suppression thresholds, etc. determine this tradeoff.



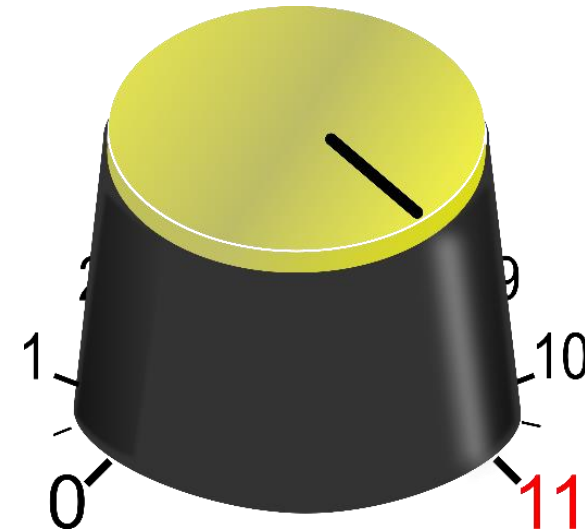


# Problem #2 – How much is enough?

Legacy disclosure avoidance methods provide little ability to quantify privacy protections.

When faced with rising disclosure risk, disclosure avoidance practitioners adjust their implementation parameters.

**BUT, this is largely a scattershot solution that over-protects some data, while often under-protecting the most vulnerable records.**



# Differential Privacy

DP is not a disclosure avoidance “method” as much as it is a framework for defining and then quantifying privacy protection.

Every individual that is reflected in a particular statistic contributes towards that statistic’s value.

Every statistic that you publish “leaks” a small amount of private information.

DP as a framework allows you to assess each individual’s contribution to the statistic, and to measure (and thus, limit) how much information about them will leak.



# Differential Privacy

When combined with noise injection, DP allows you to precisely control the amount of private information leakage in your published statistics.

- **Infinitely tunable** – parameter “dials” can be set anywhere from perfect privacy to perfect accuracy.
- **Privacy guarantee is mathematically provable and future-proof.**
- **The precise calibration of statistical noise enables optimal data accuracy for any given level of privacy protection.\***

**\*Absent post-processing requirements, which can introduce error independent of that needed to protect privacy.**



# Privacy vs. Accuracy

The only way to absolutely eliminate all risk of re-identification would be to never release any usable data.

Differential privacy allows you to quantify a precise level of “acceptable risk,” and to precisely calibrate where on the privacy/accuracy spectrum the resulting data will be.



Data Quality | Bnae Kegouqe  
Dada Qualitg | Vrkk Jzcfkdy  
Data Qaality | Dncb PrhvBl n  
Dzte Qvality | Dncb Prtnavy  
Dfha Quapyti | Tgta Ppijacy  
Tgta Qucjity | Dfha Pnjvico  
Dncb Qhulitn | Dzhe Njivaci  
Ntue Quevdto | Dzte Privhecy  
Vrkk Zuhnvy | Dada Privacg  
Bnaq Denorbe | Data Privacy

# Establishing a Privacy-loss Budget

This measure is called the “Privacy-loss Budget” (PLB) or “Epsilon.”

$\epsilon=0$  (perfect privacy) would result in completely useless data

$\epsilon=\infty$  (perfect accuracy) would result in releasing the data in fully identifiable form



Epsilon

# Comparing Methods

## Data Accuracy

Differentially private disclosure avoidance methods are not inherently better or worse than traditional methods.

Both can have varying degrees of impact on data quality depending on the parameters selected and the methods' implementation.

## Privacy

Differentially private methods are substantially better than traditional methods for protecting privacy, insofar as they actually allow for measurement of the privacy risk.

# Implications for the 2020 Census

The modernization of our privacy protections using a differential privacy framework does not change the constitutional mandate to apportion the House of Representatives according to the actual enumeration.

As in 2000 and 2010, the Census Bureau will apply privacy protections to the P.L. 94-171 redistricting data.

# Privacy-loss Budget Allocation

The Census Bureau's Data Stewardship Executive Policy Committee (DSEP) will be making decisions about the PLB for the 2020 Census. This includes allocation across different 2020 Census data products, including:

- P.L. 94-171 Redistricting data
- Demographic and Housing Characteristics files (DHC)
- Detailed Demographic and Housing Characteristics files (D-DHC)
- ...and other uses of Decennial Census data.

DSEP will also be deciding how to allocate the PLB across the different sets of tabulations *within* each data product (by geographic level and by data element).



# Recent Activity: DAS Tuning for the Redistricting Data

## P.L. 94-171 Tuning & Privacy-Accuracy Trade-off Experiments

- In December through March, the DAS Team conducted over 600 full-scale TDA runs with the complete P.L. 94-171 data product schema.
- Goal: Evaluating resulting accuracy of varying parameters for:
  - Overall setting of PLB
  - Query strategy
  - Allocation of PLB across geographic levels
  - Allocation of PLB across queries
- Worked with subject matter experts in Demographic and Decennial Directorates to evaluate accuracy of experimental runs to inform parameter setting.

# Demonstration Data

- Since October 2019, the Census Bureau has been periodically releasing demonstration data products (using 2010 Census data) for data user evaluation.
- The first four of these sets of demonstration data (October 2019, May 2020, September 2020, November 2020) used a conservative global PLB set by DSEP for the October 2019 Demonstration Product, in order to evaluate algorithmic improvements.
- ***The 2020 Census Data Products will not be held to this fixed PLB.***
- On April 28, we released another set of Privacy-Protected Microdata Files (PPMFs) and Detailed Summary Metrics using a different global PLB ( $\epsilon=12.2$ ) that more closely approximates the level of PLB that the DSEP will be considering for the 2020 Census redistricting data files.
- In September, we plan to release a final set of PPMFs using the actual production code and settings that will be used for the 2020 Census redistricting data files.

## How to Submit Feedback

The changes in the [April 2021 PPMFs](#) data set reflect the cumulative feedback received from the data user community throughout the development process. We look forward to feedback from data users on this [new demonstration product](#). Your input will inform the Census Bureau's June 2021 final decision on the PLB and on the 2020 Census redistricting data parameters. **The deadline to submit feedback is May 28, 2021.**

**\*\* Please send comments to [2020DAS@census.gov](mailto:2020DAS@census.gov) with the subject line "April 2021 Demonstration Data."**

Particularly useful feedback would describe:

• **Fitness-for-use:** Based on your analysis, would the data needed for your applications (redistricting, Voting Rights Act analysis, estimates, projections, funding data sets, etc.) be satisfactory?

- How did you come to that conclusion?
- If your analysis found the data to be unsatisfactory, how incrementally would accuracy need to change to improve the use of the data for your required or programmatic use case(s)?
- Have you identified any improbable results in the data that would be helpful for us to understand?"

• **Privacy:** Do the proposed products present any confidentiality concerns that we should address in the DAS?

• **Improvements:** Are there improvements you've identified that you want to make sure we retain in the final design? Be specific about the geography and error metric for the proposed improvement.

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

Stay Informed:  
Subscribe to the 2020 Census Data  
Products Newsletters

\*Search “Disclosure Avoidance” at [www.census.gov](http://www.census.gov)

The screenshot shows the top portion of a webpage. At the top is an orange banner with the text "2020 Census Population Counts for Apportionment are Now Available". Below this is a breadcrumb trail: "Census.gov > 2020 Census Research, Operational Plans, and Oversight > Process > Disclosure Avoidance Modernization > 2020 Census Data Products Newsletters". The main heading is "2020 Census Data Products Newsletters" with social media icons for Facebook, Twitter, and LinkedIn to its left. Below the heading is a paragraph: "Sign up for news and information about 2020 Census Data Products and the implementation of the new Disclosure Avoidance System." A prominent orange button labeled "SIGN-UP FOR NEWSLETTERS" is positioned below the text. The "Past Issues:" section follows, listing several articles with their dates and titles, separated by horizontal lines.

**2020 Census Population Counts for Apportionment are Now Available**

Census.gov > 2020 Census Research, Operational Plans, and Oversight > Process > Disclosure Avoidance Modernization > 2020 Census Data Products Newsletters

**2020 Census Data Products Newsletters**

Sign up for news and information about 2020 Census Data Products and the implementation of the new Disclosure Avoidance System.

**SIGN-UP FOR NEWSLETTERS**

**Past Issues:**

---

April 28, 2021  
**New DAS Update Meets or Exceeds Redistricting Accuracy Targets**

---

April 19, 2021  
**New Demonstration Data Will Feature Higher Privacy-loss Budget**

---

April 07, 2021  
**Meeting Redistricting Data Requirements: Accuracy Targets**

---

February 23, 2021  
**The Road Ahead: Upcoming Disclosure Avoidance System Milestones**

---

February 03, 2021  
**New DAS Phase: Optimizing Tunable Elements**


---

November 25, 2020  
**Invariants Set for 2020 Census Data Products**

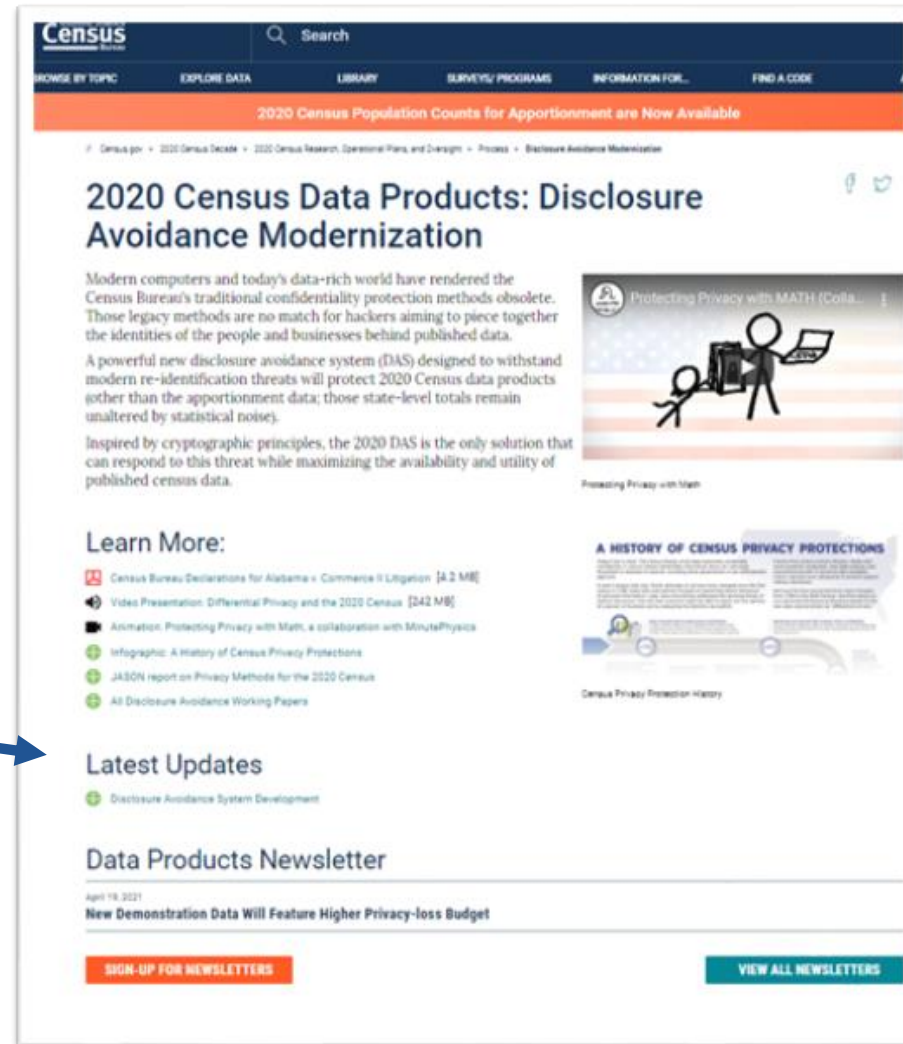
Stay Informed:  
Visit Our Website

\*Search “Disclosure Avoidance” at [www.census.gov](http://www.census.gov)

## Latest Updates

 [Disclosure Avoidance System Development](#)

Dates and log-in information for  
webinar series  
*\*coming soon*



The screenshot shows the top of the Census.gov website with a search bar and navigation menu. The main headline is "2020 Census Population Counts for Apportionment are Now Available". Below that, the article title "2020 Census Data Products: Disclosure Avoidance Modernization" is displayed. The article text discusses the obsolescence of traditional confidentiality methods and the introduction of a new Disclosure Avoidance System (DAS) designed to withstand modern re-identification threats. It mentions that the DAS is inspired by cryptographic principles and maximizes the availability and utility of published census data. To the right of the text is an illustration titled "Protecting Privacy with MATH (Colla...)" showing stick figures with a laptop and a document. Below the article text is a "Learn More:" section with a list of links: "Census Bureau Declarations for Alabama v. Commerce II Litigation [4.2 MB]", "Video Presentation: Differential Privacy and the 2020 Census [242 MB]", "Animation: Protecting Privacy with Math, a collaboration with MinutePhysics", "Infographic: A History of Census Privacy Protections", "JASON report on Privacy Methods for the 2020 Census", and "All Disclosure Avoidance Working Papers". To the right of this list is an infographic titled "A HISTORY OF CENSUS PRIVACY PROTECTIONS". Below the infographic is a "Latest Updates" section with a link for "Disclosure Avoidance System Development". At the bottom of the page is a "Data Products Newsletter" section with a date of "April 19, 2021" and a headline "New Demonstration Data Will Feature Higher Privacy-loss Budget". There are two buttons: "SIGN-UP FOR NEWSLETTERS" and "VIEW ALL NEWSLETTERS".

[START HERE >](#)

# Questions?



# Understanding the 2020 Census Disclosure Avoidance System:

## *Differential Privacy 201 and the TopDown Algorithm*

**Michael Hawes and Michael Ratcliffe**  
U.S. Census Bureau

May 13, 2021

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
**2020**

# Acknowledgements

**This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, from the following Census Bureau divisions and outside organizations:** ADCOM, ADDC, ADRM, CED, CEDDA, CEDSCI, CES, CSRM, DCMD, DITD, ESMD, GEO, POP, TAB, CDF, Econometrica Inc., Galois, Knexus Research Corp, MITRE, NLT, TI, and Tumult Labs.

**We also acknowledge and greatly appreciate the ongoing feedback we have received from external stakeholder groups that has contributed to the design and improvement of the Disclosure Avoidance System.**

**For more information and technical details relating to the issues discussed in these slides, please contact the author at [michael.b.hawes@census.gov](mailto:michael.b.hawes@census.gov).**

**Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.**

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
**2020**



# TDA System Requirements

The 2020 Disclosure Avoidance System's TopDown Algorithm (TDA) will implement formal privacy protections for the P. L. 94-171 Redistricting Data Summary File, Demographic Profiles, Demographic and Housing Characteristics, and Special Tabulations of the 2020 Census.

TDA system requirements include:

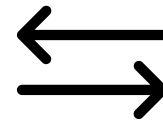
- Input/Output specifications
- Invariants
- Constraints
- Utility/Accuracy for pre-specified tabulations
- $\epsilon$ -asymptotic consistency
- Transparency

# TDA Process Snapshot



# What is a histogram?

Record ID	Block	Race	...	Sex
1	1001	Black	...	Male
2	1001	Black	...	Male
3	1001	Asian	...	Female
4	1001	Asian	...	Female
5	1001	Black	...	Male
6	1001	AIAN	...	Female
7	1001	AIAN	...	Male
8	1001	Black	...	Female
9	1001	Black	...	Female



Attribute Combination (Block/Race/.../Sex)	# of Records
1001/AIAN/.../Male	1
1001/AIAN/.../Female	1
1001/Asian/.../Male	0
1001/Asian/.../Female	2
1001/Black/.../Male	3
1001/Black/.../Female	2
...	...

Histogram: Record count for each unique combination of attributes (including location)

Microdata: One record per respondent

# Noisy Measurements

**TDA allocates shares of the total privacy-loss budget by geographic level and by query.**

**Each query of the confidential data will have noise added to its answer.**

**The noise is taken from a probability distribution with mean=0, and variance determined by the share of the PLB allocated to that particular query at that geographic level.**

**These noisy measurements are independent of each other, and can include negative values, hence the need for post-processing.**



# What is noise?

To protect privacy, TDA randomly adds or subtracts a small amount from each statistic it calculates from the confidential data.

Attribute Combination (Block/Race/.../Sex)	# of Records
1001/AIAN/.../Male	1
1001/AIAN/.../Female	1
1001/Asian/.../Male	0
1001/Asian/.../Female	2
1001/Black/.../Male	3
1001/Black/.../Female	2
...	...

# Total:  $9+0=9$

# Male:  $4+0=4$

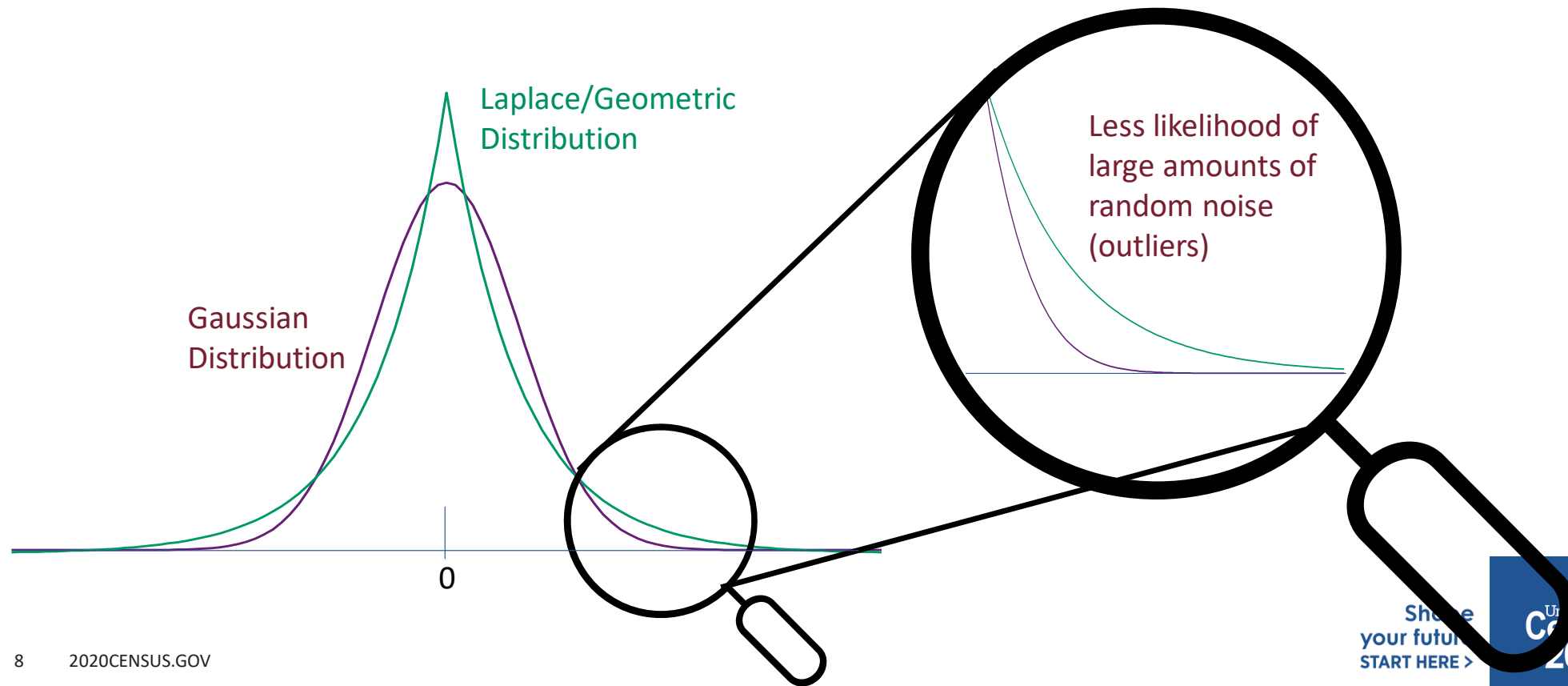
# Female:  $5-1=4$

#AIAN:  $2+0=2$

#Asian:  $2+2=4$

#Black:  $5-1=4$

# Zero-Concentrated Differential Privacy (zCDP)



# Understanding *epsilon*, *delta* and *rho*

## In traditional ( $\epsilon, 0$ ) differential privacy:

The privacy-loss parameter  $\epsilon$  (*epsilon*) sets the upper-bound on how much information leakage can occur.

Shares of  $\epsilon$  are allocated to each query and sum to the global value of  $\epsilon$ .

## In zero-concentrated differential privacy (zCDP):

Privacy loss is quantified by the paired parameters  $\epsilon$  and  $\delta$  (*delta*).

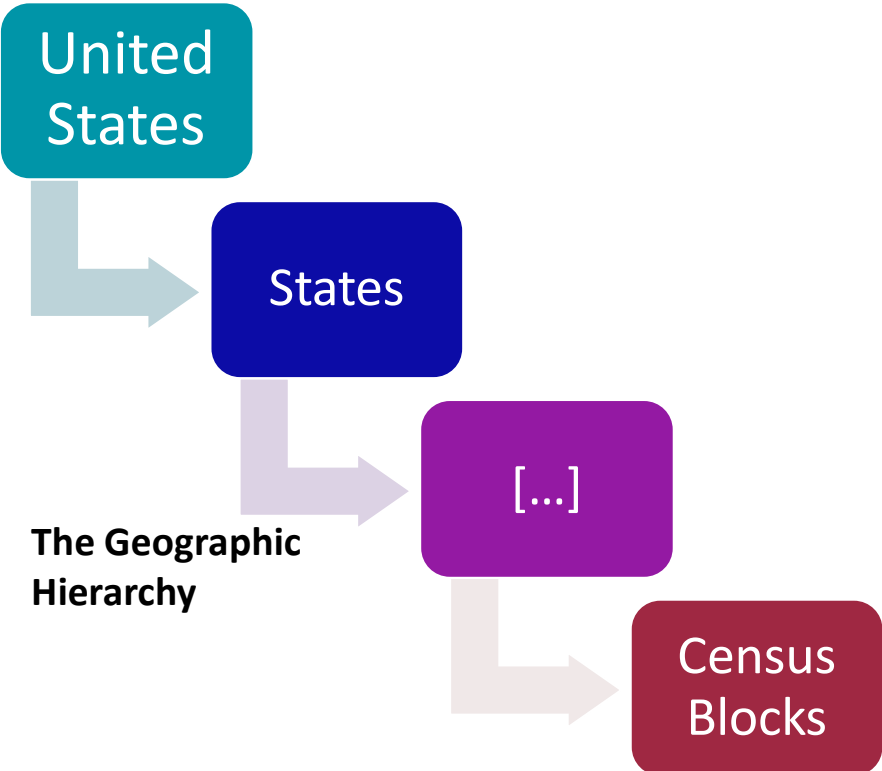
$\delta$  is a probabilistic term that establishes the likelihood that privacy loss might exceed the upper bound represented by a particular value of  $\epsilon$ .

Within the mechanics of zCDP, privacy-loss budget is allocated to queries by shares of a third parameter,  $\rho$  (*rho*).

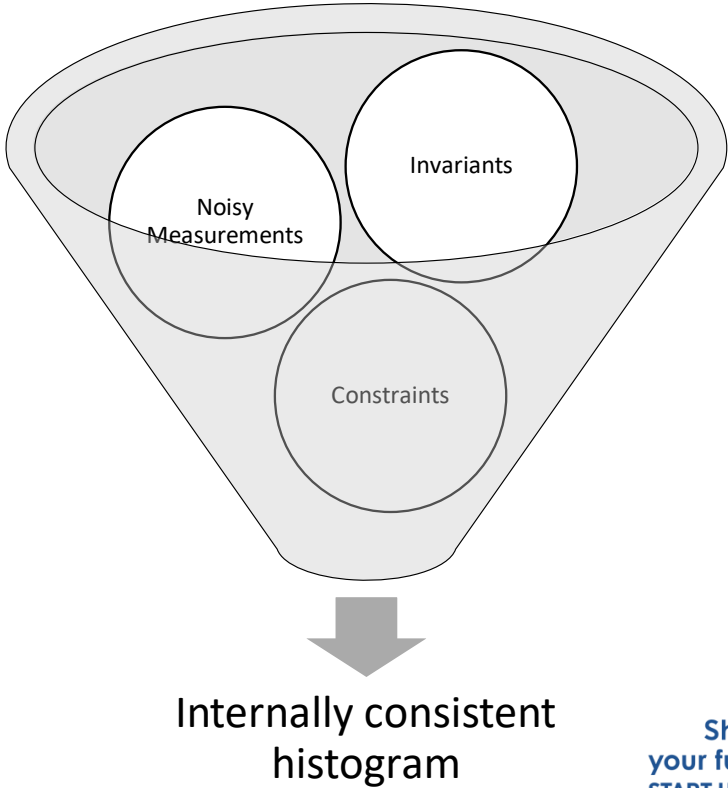
The global  $\rho$  can then be used to calculate the global  $\epsilon$  for any given level of  $\delta$ .

The Census Bureau's privacy accounting uses a value of  $\delta=10^{-10}$  so our published values of  $\epsilon$  should be interpreted accordingly.

# The TopDown Approach



At each geographic level:

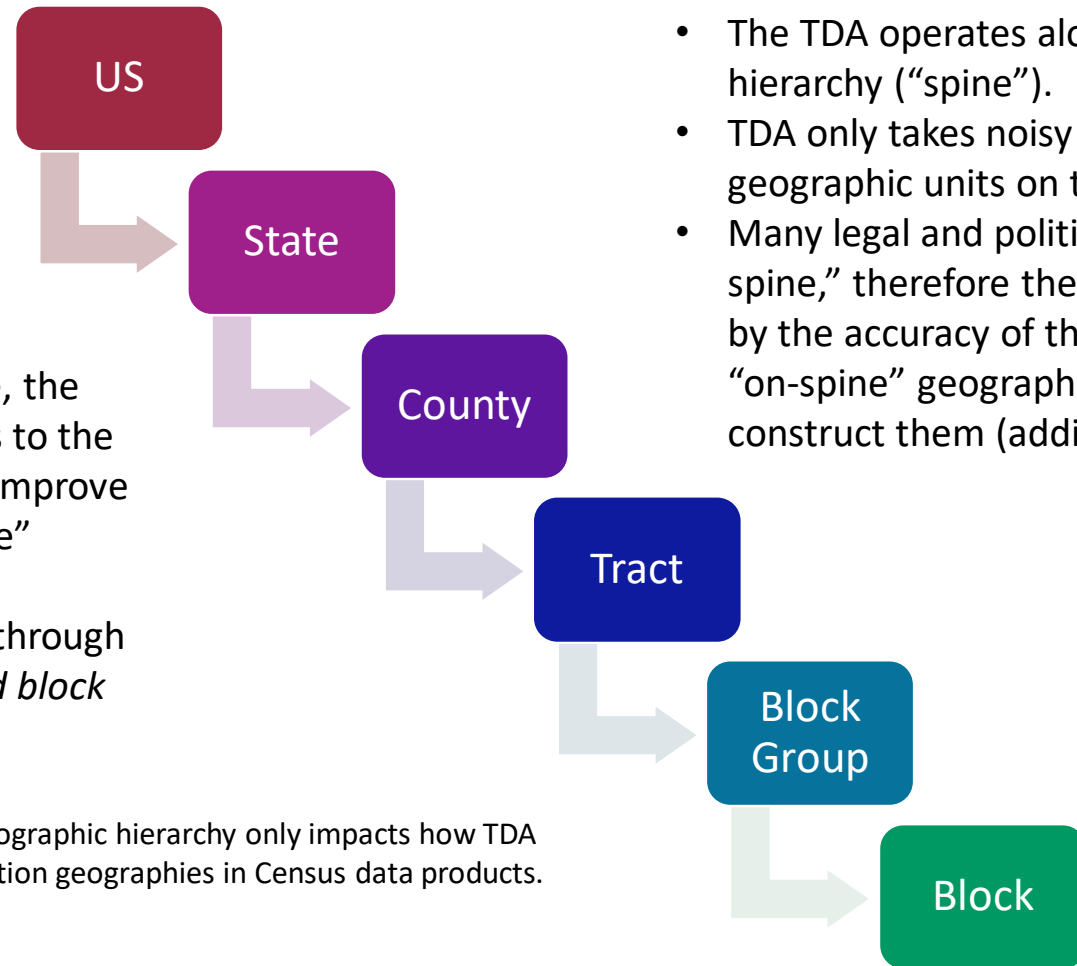




# Benefits of TDA Compared to Block-by-block

- TDA is in stark contrast with naïve alternatives (e.g., block-by-block or bottom-up)
- TDA disclosure-limitation error does not increase with number of contained Census blocks in the geographic entity
- TDA yields increasing relative accuracy as the population being measured increases (in general), and increased count accuracy compared to block-by-block
- TDA “borrows strength” from upper geographic levels to improve count accuracy at lower geographic levels (e.g., for sparsity)

# Tabulation Geographic Hierarchy



- To address this challenge, the DAS Team made changes to the geographic hierarchy to improve the accuracy of “off-spine” geographies.
- This was done primarily through the creation of *optimized block groups (not shown)*.

- The TDA operates along a geographic hierarchy (“spine”).
- TDA only takes noisy measurements for geographic units on the hierarchy.
- Many legal and political geographies are “off-spine,” therefore their accuracy is impacted by the accuracy of the minimum number of “on-spine” geographies that can be used to construct them (adding or subtracting).

Note: The optimization of the geographic hierarchy only impacts how TDA operates. It will not affect tabulation geographies in Census data products.

# Rethinking the Geographic Hierarchy

## Geographic Hierarchy for Disclosure Avoidance System Processing

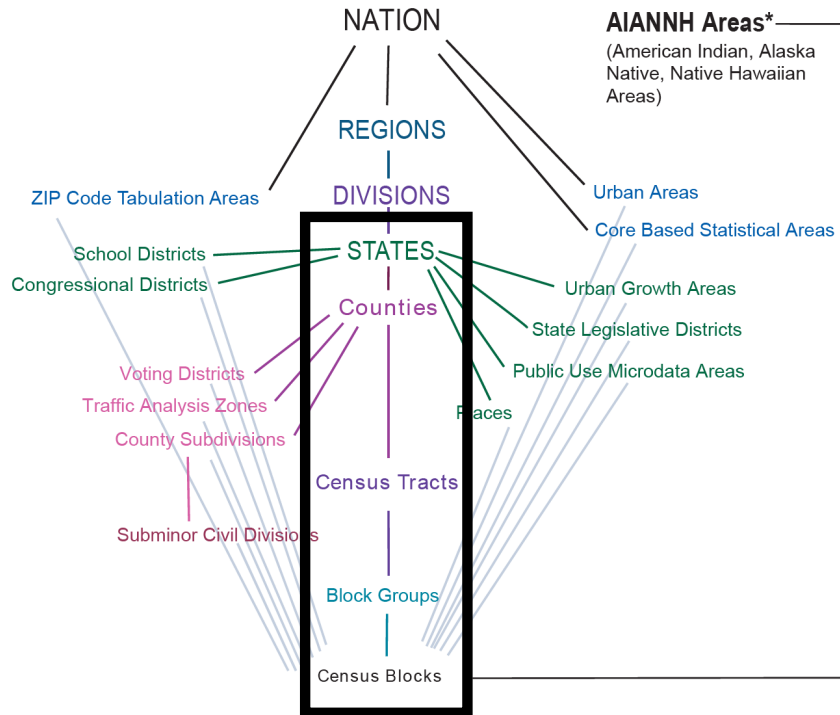
*Challenge:* Provide for the direct measurement of population and characteristics for American Indian/Alaska Native/Native Hawaiian (AIANNH) areas and sub-state legal geography when applying differential privacy methods.

*Consideration:* The larger the number of geographic areas on the geographic hierarchy (“spine”) and the more intersections between geographic areas that are formed when one type of area overlaps with another, the more thinly the privacy-loss budget is distributed, impacting the accuracy of data for all geographic areas.

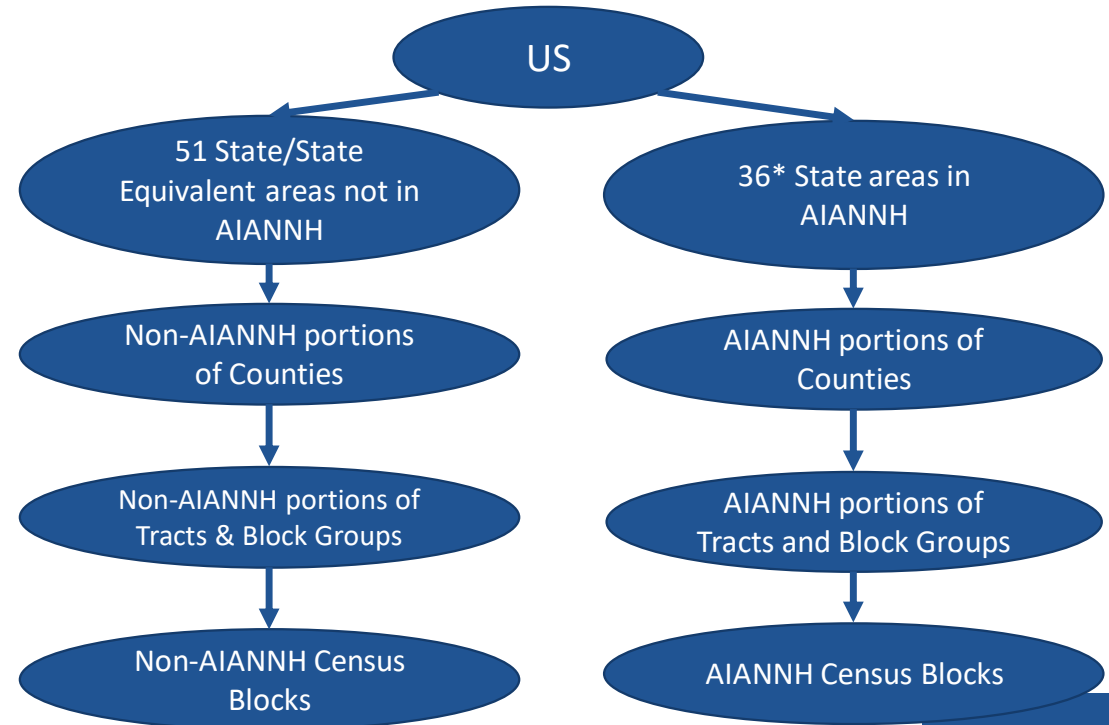
*Solution:* Bring legal AIANNH areas as well as places (incorporated places and census designated places in 38 states; cities and towns/townships in 12 states) closer to the spine for Disclosure Avoidance System (DAS) processing.

# Revising the geographical hierarchy for disclosure avoidance processing

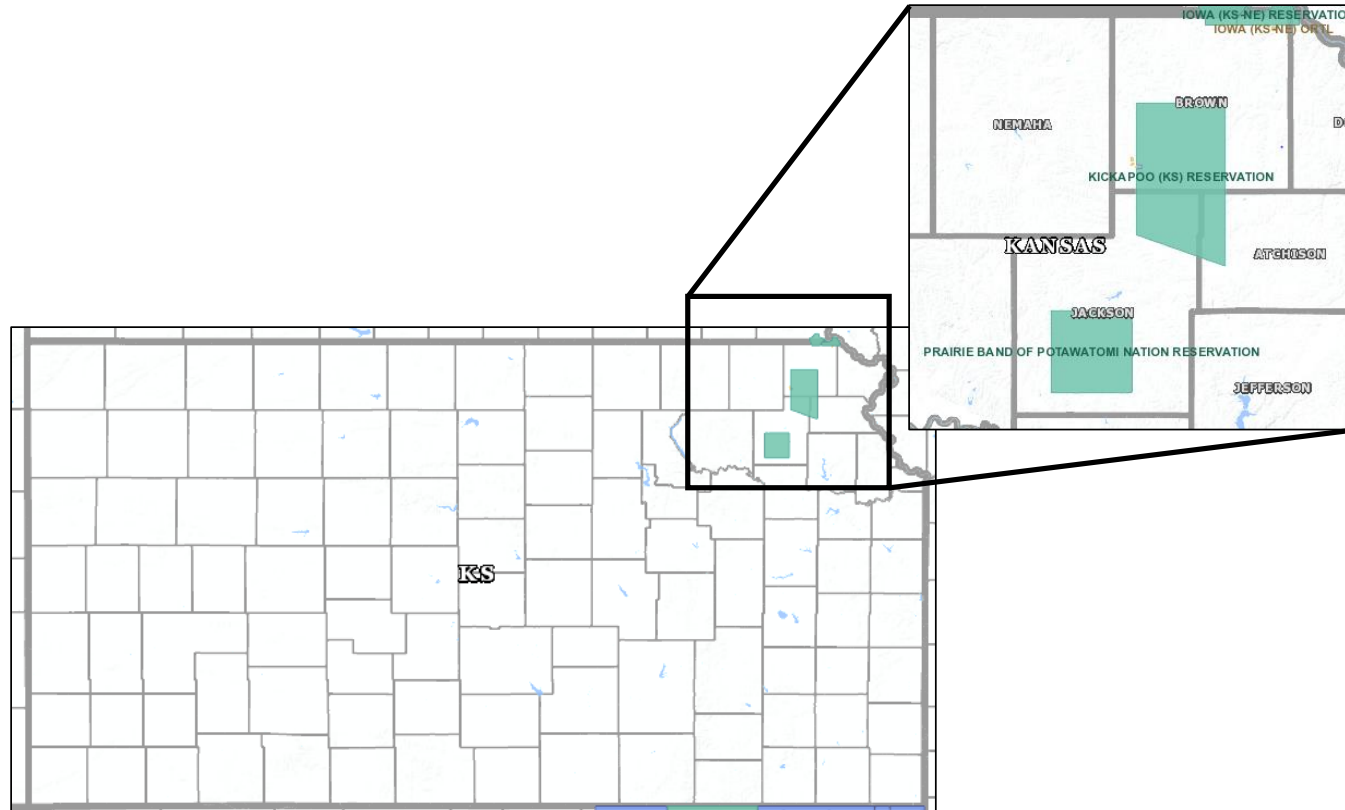
## Standard Hierarchy:



## Hierarchy for DAS Processing (high-level):



# Providing for Direct Measurement of American Indian, Alaska Native, and Native Hawaiian Areas

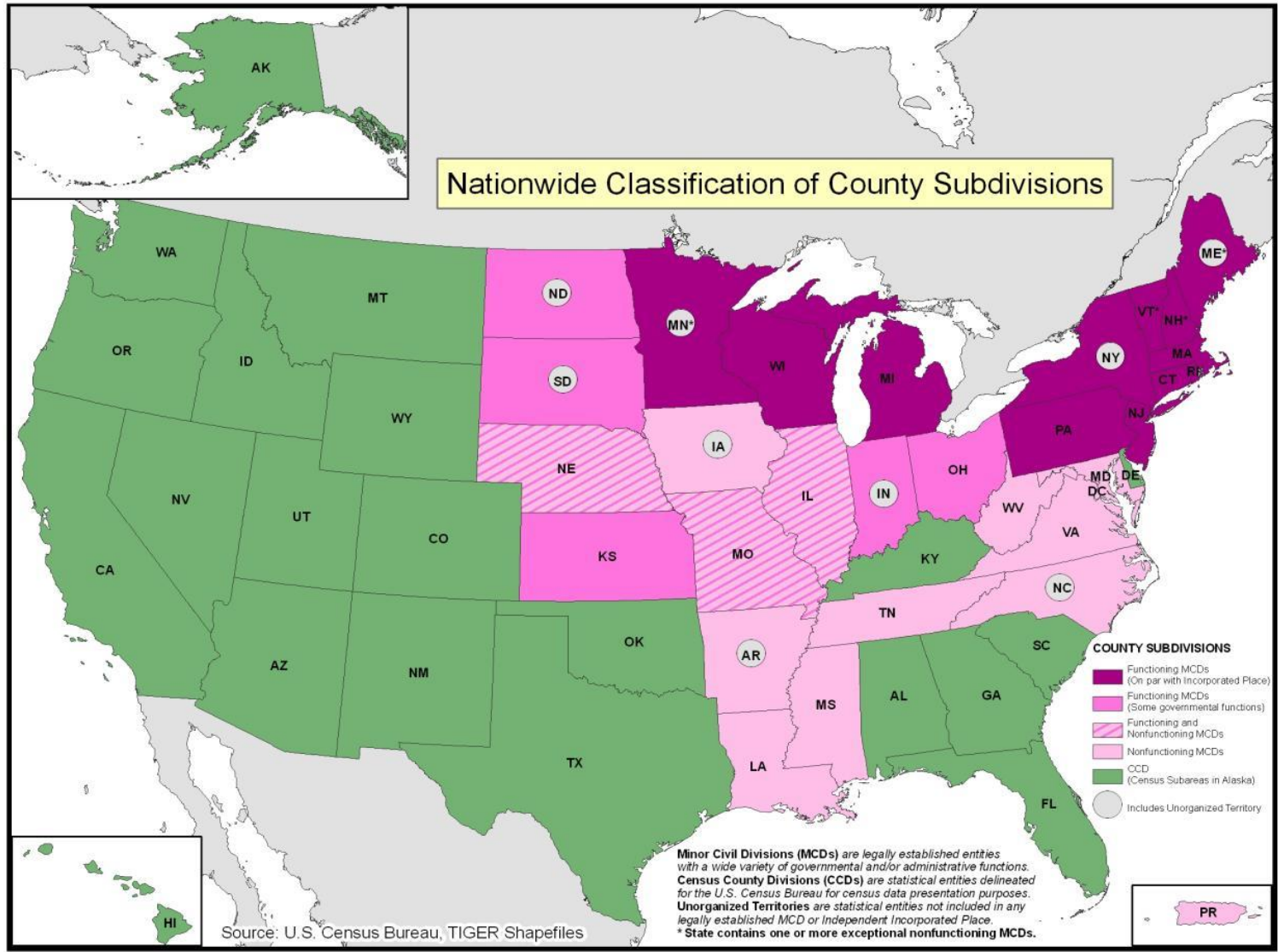


**All AIANNH Areas within the state as a single group, providing a population count for all areas within a state. This minimizes the likelihood that post-processing could result in systematic undercounts.**

## **Example:**

The three American Indian areas in Kansas grouped together at the “state” level:

- Iowa (KS-NE) Reservation and Off-Reservation Trust Lands +
- Kickapoo (KS) Reservation +
- Prairie Band of Potawatomi Nation Reservation.



The 12 “Strong-Minor Civil Division” (MCD) states are those in purple.

The MCDs (cities, boroughs, and towns/townships) in these states have active functioning governments on par with incorporated places in other states.

Shape  
your future  
START HERE >

United States  
**Census**  
2020

Focusing the geographic hierarchy on the more important sub-state geographic entities in recognition of the regional variations that exist.

**Optimized Block Groups (high-level):**

**In the 38 “non-strong-Minor Civil Division” States, District of Columbia, and Puerto Rico:**

**Optimized Block Groups were configured to bring Places (Summary Level 160) closer to the spine.**



**In the 12 “Strong-Minor Civil Division” States:**

**Optimized Block Groups were configured to bring Minor Civil Divisions (e.g., cities, boroughs, and towns/townships) closer to the spine.**





# Multi-pass Post-processing

The sparsity of many queries (i.e., prevalence of zeros and small counts) has the potential to introduce bias in TDA's post-processing.

To address the sparsity issue, TDA processing is now performed in a series of passes.

At certain geographic levels, the algorithm constructs histograms for a subset of queries in a series of passes for that level, constraining the histogram for each pass to be consistent with the histogram produced in the prior pass.

Example for the P.L. 94-171 Redistricting Data Summary File:

Pass 1: Total Population

Pass 2: Remaining tabulations supporting P.L. 94-171 Redistricting Data

# Sample Privacy-loss Budget Allocation (by geographic level)

Privacy-loss Budget Allocation April 28, 2021	
PPMF	
Person Tables (PPMF-P)	
United States	
Global rho	192721/184041 (1.05)
Global epsilon	10.3
delta	10 <sup>-10</sup>
rho Allocation by Geographic Level	
US	51/1024
State	153/1024
County	78/1024
Tract	51/1024
Optimized Block Group*	172/1024
Block	519/1024

Privacy-loss Budget Allocation April 28, 2021	
PPMF	
Units Tables (PPMF-U)	
United States	
Global rho	919681/20241001 (0.045)
Global epsilon	1.9
delta	10 <sup>-10</sup>
rho Allocation by Geographic Level	
US	1/1024
State	1/1024
County	18/1024
Tract	75/1024
Optimized Block Group*	906/1024
Block	23/1024

## Sample Privacy-loss Budget Allocation (by query)

Query	Per Query rho Allocation by Geographic Level					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		678/1024**	342/1024	1/1024	572/1024	1/1024
CENRACE (63 cells)	2/1024	1/1024	1/1024	2/1024	1/1024	2/1024
HISPANIC (2 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
VOTINGAGE (2 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
HHINSTLEVELS (3 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
HHGQ (8 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
HISPANIC*CENRACE (126 cells)	5/1024	2/1024	3/1024	5/1024	3/1024	5/1024
VOTINGAGE*CENRACE (126 cells)	5/1024	2/1024	3/1024	5/1024	3/1024	5/1024
VOTINGAGE*HISPANIC (4 cells)	1/1024	1/1024	1/1024	1/1024	1/1024	1/1024
VOTINGAGE*HISPANIC*CENRACE (252 cells)	17/1024	6/1024	11/1024	17/1024	8/1024	17/1024
HHGQ*VOTINGAGE*						
HISPANIC*CENRACE (2,016 cells)	990/1024	330/1024	659/1024	989/1024	432/1024	989/1024

\*The Optimized Block Groups used within the TopDown Algorithm differ from tabulation block groups. These differences improve accuracy for "off-spine" geographies like places and minor civil divisions. The use of optimized block groups for measurement and post-processing within the TopDown Algorithm does not impact how the resulting data will be tabulated. All Census data products will be tabulated using the official tabulation block groups as defined by the Census Bureau's Geography Division.

\*\*The TOTAL query (total population) is held invariant at the state level. This rho allocation assigned to TOTAL at the state level is the amount assigned to the state-level queries for the total population of all American Indian and Alaska Native (AIAN) tribal areas within the state and for the total population of the remainder of the state, for the 36 states that include AIAN tribal areas.

Webinar Series:

## Understanding the 2020 Census Disclosure Avoidance System

All webinars start at **1:00 pm EDT**

No pre-registration necessary.

\*Search “*disclosure webinars*” at [www.census.gov](http://www.census.gov) for log-in information and archived presentations.

Or go to: <https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series.html>

Day	Date	Title
T	May 4	Differential Privacy 101
F	May 7	The Census Bureau's Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census
Th	May 13	Differential Privacy 201 and the TopDown Algorithm
F	May 14	Highlights of the April 2021 Detailed Summary Metrics
F	May 21	Analysis of April 2021 Demonstration Data for Redistricting and Voting Rights Act Use Cases

2020CENSUS.GOV

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

IRC\_01157

Stay Informed:  
Subscribe to the 2020 Census Data  
Products Newsletters

\*Search “Disclosure Avoidance” at [www.census.gov](http://www.census.gov)



## 2020 Census Data Products Newsletters

Sign up for news and information about 2020 Census Data Products and the implementation of the new Disclosure Avoidance System.

[SIGN-UP FOR NEWSLETTERS](#)

### Past Issues:

May 04, 2021

**Webinar Today (5/4): Differential Privacy 101**

April 30, 2021

**Save the Dates for Additional Webinars Throughout May**

April 28, 2021

**New DAS Update Meets or Exceeds Redistricting Accuracy Targets**

April 19, 2021

**New Demonstration Data Will Feature Higher Privacy-loss Budget**

April 07, 2021

**Meeting Redistricting Data Requirements: Accuracy Targets**

February 23, 2021

**The Road Ahead: Upcoming Disclosure Avoidance System Milestones**

[START HERE >](#)

Stay Informed:  
Visit Our Website

\*Search “Disclosure Avoidance” at [www.census.gov](http://www.census.gov)

***“Disclosure Avoidance Webinar Series:  
Join live or view archived presentations”***



## 2020 Census Data Products: Disclosure Avoidance Modernization

Modern computers and today's data-rich world have rendered the Census Bureau's traditional confidentiality protection methods obsolete. Those legacy methods are no match for hackers aiming to piece together the identities of the people and businesses behind published data.

A powerful new disclosure avoidance system (DAS) designed to withstand modern re-identification threats will protect 2020 Census data products (other than the apportionment data; those state-level totals remain unaltered by statistical noise).

Inspired by cryptographic principles, the 2020 DAS is the only solution that can respond to this threat while maximizing the availability and utility of published census data.

### Learn More:

- \*\* Disclosure Avoidance Webinar Series: Join live or view archived presentations \*\*
- Census Bureau Declarations for Alabama v. Commerce II Litigation [4.2 MB]
- Video Presentation: Differential Privacy and the 2020 Census [242 MB]
- Animation: Protecting Privacy with Math, a collaboration with MinutePhysics
- Infographic: A History of Census Privacy Protections
- JASON report on Privacy Methods for the 2020 Census
- All Disclosure Avoidance Working Papers

### Latest Updates

- Disclosure Avoidance System Development

### Data Products Newsletter

April 30, 2021  
Save the Dates for Additional Webinars Throughout May

[SIGN-UP FOR NEWSLETTERS](#) [VIEW ALL NEWSLETTERS](#)

## EMPIRICAL STUDY of TWO ASPECTS of THE

**IRC\_01160**

1/37

**EMPIRICAL STUDY of TWO ASPECTS of THE  
TOPDOWN ALGORITHM OUTPUT for REDISTRICTING:**



**EMPIRICAL STUDY of TWO ASPECTS of THE  
TOPDOWN ALGORITHM OUTPUT for REDISTRICTING:  
RELIABILITY & VARIABILITY**

**EMPIRICAL STUDY of TWO ASPECTS of THE  
TOPDOWN ALGORITHM OUTPUT for REDISTRICTING:  
RELIABILITY & VARIABILITY**

*PRE-DECISIONAL: For Internal & DOJ Comments Only*

2021-05-18

Tommy Wright and Kyle Irimata  
Center for Statistical Research and Methodology  
Research and Methodology Directorate  
U.S. Bureau of the Census  
Washington, D.C. 20233

**IRC\_01163**

1/37

---

*Disclaimer and Acknowledgements:* The views presented in this paper are those of the authors and not the U.S. Bureau of the Census.

---

*Disclaimer and Acknowledgements:* The views presented in this paper are those of the authors and not the U.S. Bureau of the Census. We are grateful to our colleagues Mary Mulry, Pat Cantwell, Eric Slud, and James Whitehorne for their reading of a draft of this study and for their comments and questions that have strengthened the paper's presentation and content. We are also grateful for the many conversations with members of the Disclosure Avoidance System (DAS) Team. The statistics in this paper have been cleared by the Census Bureau Disclosure Review Board (DRB Clearance Number CBDRB-FY20-DSEP-001).  
*Corresponding author for comments:* [tommy.wright@census.gov](mailto:tommy.wright@census.gov); (301) 763-1702.

## TECHNICAL SUMMARY

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

QUESTION:

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

QUESTION: *“What is the minimum TOTAL (ideal) population of a district*



## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

QUESTION: *“What is the minimum TOTAL (ideal) population of a district to have reliable characteristics*

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

QUESTION: *"What is the minimum TOTAL (ideal) population of a district to have reliable characteristics of various demographic groups?"*

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

QUESTION: *"What is the minimum TOTAL (ideal) population of a district to have reliable characteristics of various demographic groups?"*

For each of nearly 200,000 block groups (proxies for voting districts) in the United States,

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

QUESTION: *"What is the minimum TOTAL (ideal) population of a district to have reliable characteristics of various demographic groups?"*

For each of nearly 200,000 block groups (proxies for voting districts) in the United States,

ANSWER:

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

QUESTION: *“What is the minimum TOTAL (ideal) population of a district to have reliable characteristics of various demographic groups?”*

For each of nearly 200,000 block groups (proxies for voting districts) in the United States,

ANSWER: *“for any block group with a TOTAL count near 600 people,*

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

QUESTION: *“What is the minimum TOTAL (ideal) population of a district to have reliable characteristics of various demographic groups?”*

For each of nearly 200,000 block groups (proxies for voting districts) in the United States,

ANSWER: *“for any block group with a TOTAL count near 600 people, the difference between the TDA ratio of the largest demographic group (LDG)*

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

*QUESTION: "What is the minimum TOTAL (ideal) population of a district to have reliable characteristics of various demographic groups?"*

For each of nearly 200,000 block groups (proxies for voting districts) in the United States,

*ANSWER: "for any block group with a TOTAL count near 600 people, the difference between the TDA ratio of the largest demographic group (LDG) and the corresponding SWA ratio for the LDG"*

## TECHNICAL SUMMARY

In Part I of this limited study ( $\epsilon = 10.3$ , for person file),

*QUESTION: "What is the minimum TOTAL (ideal) population of a district to have reliable characteristics of various demographic groups?"*

For each of nearly 200,000 block groups (proxies for voting districts) in the United States,

*ANSWER: "for any block group with a TOTAL count near 600 people, the difference between the TDA ratio of the largest demographic group (LDG) and the corresponding SWA ratio for the LDG is less than or equal to 5 percentage points at least 95% of the time".*



- We also consider “places and minor civil divisions (MCDs)” as proxies for districts.

- We also consider “places and minor civil divisions (MCDs)” as proxies for districts. A similar minimum TOTAL between 350 and 400 is observed for places and MCDs.

- We also consider “places and minor civil divisions (MCDs)” as proxies for districts. A similar minimum TOTAL between 350 and 400 is observed for places and MCDs.
- No congressional or state legislative district

- We also consider “places and minor civil divisions (MCDs)” as proxies for districts. A similar minimum TOTAL between 350 and 400 is observed for places and MCDs.
- No congressional or state legislative district failed our test for reliability.

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE:

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA*

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island



Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS:

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS: *Given more development of the TDA,*

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS: *Given more development of the TDA, a larger  $\epsilon$ ,*

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS: *Given more development of the TDA, a larger  $\epsilon$ , and additional focus on how to allocate this  $\epsilon$ ,*

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS: *Given more development of the TDA, a larger  $\epsilon$ , and additional focus on how to allocate this  $\epsilon$ , we see less variability throughout with output from the latest TDA.*

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS: *Given more development of the TDA, a larger  $\epsilon$ , and additional focus on how to allocate this  $\epsilon$ , we see less variability throughout with output from the latest TDA.*

FINDINGS:

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS: *Given more development of the TDA, a larger  $\epsilon$ , and additional focus on how to allocate this  $\epsilon$ , we see less variability throughout with output from the latest TDA.*

FINDINGS: As we reported in [5],



Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS: *Given more development of the TDA, a larger  $\epsilon$ , and additional focus on how to allocate this  $\epsilon$ , we see less variability throughout with output from the latest TDA.*

FINDINGS: As we reported in [5], relative variability in the *TDA* increases

Part II (previous  $\epsilon = 4.0$ ; latest  $\epsilon = 10.3$ )

OBJECTIVE: Assess the variability of the 2021-04-28 version of the *TDA* for congressional districts and state legislative districts in Rhode Island and for three additional jurisdictions shared by the U.S. Department of Justice.

FINDINGS: *Given more development of the TDA, a larger  $\epsilon$ , and additional focus on how to allocate this  $\epsilon$ , we see less variability throughout with output from the latest TDA.*

FINDINGS: As we reported in [5], relative variability in the *TDA* increases as we consider smaller pieces of geography and population.

# Part I

## I.1. INTRODUCTION

## Part I

### I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district*

## Part I

### I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

## Part I

### I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States,

## Part I

### I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States, we compare closeness between:

## Part I

### I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States, we compare closeness between:

- (a) published SWA counts



# Part I

## I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States, we compare closeness between:

- (a) published SWA counts based on a Swapping Algorithm (SWA)

## Part I

### I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States, we compare closeness between:

(a) published SWA counts based on a Swapping Algorithm (SWA) applied to the 2010 Census Edited File and

# Part I

## I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States, we compare closeness between:

- (a) published *SWA* counts based on a Swapping Algorithm (*SWA*) applied to the 2010 Census Edited File and
- (b) the corresponding *TDA* counts

# Part I

## I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States, we compare closeness between:

- (a) published *SWA* counts based on a Swapping Algorithm (*SWA*) applied to the 2010 Census Edited File and
- (b) the corresponding *TDA* counts based on the 2021-04-28 version of the *TDA*

# Part I

## I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States, we compare closeness between:

- (a) published *SWA* counts based on a Swapping Algorithm (*SWA*) applied to the 2010 Census Edited File and
- (b) the corresponding *TDA* counts based on the 2021-04-28 version of the *TDA* applied to the 2010 Census Edited File.

# Part I

## I.1. INTRODUCTION

QUESTION: *“What is the minimum TOTAL (ideal<sup>a</sup>) population of a district to have reliable characteristics of various demographic groups?”*

For each of the 217,740 block groups in the United States, we compare closeness between:

- (a) published *SWA* counts based on a Swapping Algorithm (*SWA*) applied to the 2010 Census Edited File and
- (b) the corresponding *TDA* counts based on the 2021-04-28 version of the *TDA* applied to the 2010 Census Edited File.

Our comparisons are facilitated by the **difference of ratios** *DR*.

*Definition 1:*

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts



*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

*Definition 1:*

- (1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.
- (2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

Small values of the difference of ratios  $DR_g$

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

Small values of the difference of ratios  $DR_g$  imply that the ratios for a group  $g$

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

Small values of the difference of ratios  $DR_g$  imply that the ratios for a group  $g$  due to *SWA* and *TDA* in the block group

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

Small values of the difference of ratios  $DR_g$  imply that the ratios for a group  $g$  due to *SWA* and *TDA* in the block group are close.



*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

Small values of the difference of ratios  $DR_g$  imply that the ratios for a group  $g$  due to *SWA* and *TDA* in the block group are close.

*Definition 2:*

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

Small values of the difference of ratios  $DR_g$  imply that the ratios for a group  $g$  due to *SWA* and *TDA* in the block group are close.

*Definition 2:*

When  $DR_g$  is sufficiently small,

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

Small values of the difference of ratios  $DR_g$  imply that the ratios for a group  $g$  due to *SWA* and *TDA* in the block group are close.

*Definition 2:*

When  $DR_g$  is sufficiently small, we say that the  $C_{TDA}(g)$  count (or ratio)

*Definition 1:*

(1)  $C_{SWA}(g)$  and  $C_{TDA}(g)$  competing counts of demographic group  $g$  associated with a block group.

(2) Total block group counts are  $C_{SWA}$  and  $C_{TDA}$ .

The **difference of ratios** is

$$DR_g = \left| \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right|. \quad (1)$$

Small values of the difference of ratios  $DR_g$  imply that the ratios for a group  $g$  due to *SWA* and *TDA* in the block group are close.

*Definition 2:*

When  $DR_g$  is sufficiently small, we say that the  $C_{TDA}(g)$  count (or ratio) provides a **reliable characteristic** for the block group.

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560		

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	



**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198		

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	
TOTALHISP	133		

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	
TOTALHISP	133	139	

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	
TOTALHISP	133	139	$\left  \frac{133}{1,560} - \frac{139}{1,587} \right  = 0.0023$

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the TDA.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	
TOTALHISP	133	139	$\left  \frac{133}{1,560} - \frac{139}{1,587} \right  = 0.0023$
TOTALNH	1,427	1,448	$\left  \frac{1,427}{1,560} - \frac{1,448}{1,587} \right  = 0.0023$

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the TDA.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	
TOTALHISP	133	139	$\left  \frac{133}{1,560} - \frac{139}{1,587} \right  = 0.0023$
TOTALNH	1,427	1,448	$\left  \frac{1,427}{1,560} - \frac{1,448}{1,587} \right  = 0.0023$
WHITENH	1,169	1,185	$\left  \frac{1,169}{1,560} - \frac{1,185}{1,587} \right  = 0.0027$

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	
TOTALHISP	133	139	$\left  \frac{133}{1,560} - \frac{139}{1,587} \right  = 0.0023$
TOTALNH	1,427	1,448	$\left  \frac{1,427}{1,560} - \frac{1,448}{1,587} \right  = 0.0023$
WHITENH	1,169	1,185	$\left  \frac{1,169}{1,560} - \frac{1,185}{1,587} \right  = 0.0027$
BLACKNH	36	61	$\left  \frac{36}{1,560} - \frac{61}{1,587} \right  = 0.0154$



**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	
TOTALHISP	133	139	$\left  \frac{133}{1,560} - \frac{139}{1,587} \right  = 0.0023$
TOTALNH	1,427	1,448	$\left  \frac{1,427}{1,560} - \frac{1,448}{1,587} \right  = 0.0023$
WHITENH	1,169	1,185	$\left  \frac{1,169}{1,560} - \frac{1,185}{1,587} \right  = 0.0027$
BLACKNH	36	61	$\left  \frac{36}{1,560} - \frac{61}{1,587} \right  = 0.0154$
AIANNH	10	9	$\left  \frac{10}{1,560} - \frac{9}{1,587} \right  = 0.0007$
ASIANNH	187	182	$\left  \frac{187}{1,560} - \frac{182}{1,587} \right  = 0.0052$
HPINH	5	1	$\left  \frac{5}{1,560} - \frac{1}{1,587} \right  = 0.0026$
OTHERNH	11	1	$\left  \frac{11}{1,560} - \frac{1}{1,587} \right  = 0.0064$
MLTMNNH	9	9	$\left  \frac{9}{1,560} - \frac{9}{1,587} \right  = 0.0001$

**Table 1a: Block Group 240317044041 (564 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the *TDA*.)

Demographic Group ( $g$ ) <sup>b</sup>	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	1,560	1,587	
TOTAL18	1,198	1,209	
TOTALHISP	133	139	$\left  \frac{133}{1,560} - \frac{139}{1,587} \right  = 0.0023$
TOTALNH	1,427	1,448	$\left  \frac{1,427}{1,560} - \frac{1,448}{1,587} \right  = 0.0023$
WHITENH	1,169	1,185	$\left  \frac{1,169}{1,560} - \frac{1,185}{1,587} \right  = 0.0027$
BLACKNH	36	61	$\left  \frac{36}{1,560} - \frac{61}{1,587} \right  = 0.0154$
AIANNH	10	9	$\left  \frac{10}{1,560} - \frac{9}{1,587} \right  = 0.0007$
ASIANNH	187	182	$\left  \frac{187}{1,560} - \frac{182}{1,587} \right  = 0.0052$
HPINH	5	1	$\left  \frac{5}{1,560} - \frac{1}{1,587} \right  = 0.0026$
OTHERNH	11	1	$\left  \frac{11}{1,560} - \frac{1}{1,587} \right  = 0.0064$
MLTMNNH	9	9	$\left  \frac{9}{1,560} - \frac{9}{1,587} \right  = 0.0001$
HISP18	93	92	$\left  \frac{93}{1,198} - \frac{92}{1,209} \right  = 0.0015$
NONHISP18	1,105	1,117	$\left  \frac{1,105}{1,198} - \frac{1,117}{1,209} \right  = 0.0015$
WHITENH18	914	919	$\left  \frac{914}{1,198} - \frac{919}{1,209} \right  = 0.0028$
BLACKNH18	29	42	$\left  \frac{29}{1,198} - \frac{42}{1,209} \right  = 0.0105$
AIANNH18	8	9	$\left  \frac{8}{1,198} - \frac{9}{1,209} \right  = 0.0008$
ASIANNH18	142	140	$\left  \frac{142}{1,198} - \frac{140}{1,209} \right  = 0.0027$
HPINH18	2	1	$\left  \frac{2}{1,198} - \frac{1}{1,209} \right  = 0.0008$
OTHERNH18	6	1	$\left  \frac{6}{1,198} - \frac{1}{1,209} \right  = 0.0042$
MLTMNNH18	4	5	$\left  \frac{4}{1,198} - \frac{5}{1,209} \right  = 0.0008$

**Table 1b: Block Group 110010047012 (1,709 HUs) Characteristics**  
 ( $C_{TDA}(g)$  counts result from 2021-04-28 version of the TDA.)

Demographic Group ( $g$ )	$C_{SWA}(g)$	$C_{TDA}(g)$	$DR_g = \left  \frac{C_{SWA}(g)}{C_{SWA}} - \frac{C_{TDA}(g)}{C_{TDA}} \right $
TOTAL	2,875	2,902	
TOTAL18	2,261	2,280	
TOTALHISP	92	116	0.0080
TOTALNH	2,783	2,786	0.0080
WHITENH	541	529	0.0059
BLACKNH	1,686	1,697	0.0017
AIANNH	12	3	0.0031
ASIANNH	515	522	0.0007
HPINH	1	1	0.0000
OTHERNH	3	6	0.0010
MLTMNNH	25	28	0.0010
HISP18	86	100	0.0058
NONHISP18	2,175	2,180	0.0058
WHITENH18	529	519	0.0063
BLACKNH18	1,151	1,167	0.0028
AIANNH18	12	3	0.0040
ASIANNH18	460	465	0.0005
HPINH18	1	1	0.0000
OTHERNH18	3	6	0.0013
MLTMNNH18	19	19	0.0001

**CHARACTERISTICS  
of TWELVE MORE BLOCK GROUPS**

**IRC\_01237**

10/37

Demographic Group (g)	Block Group 483019501001 (TX) <sup>c</sup>			Block Group 010599729001 (AL)			Block Group 010059507002 (AL)			Block Group 040030008001 (AZ)		
	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>
<b>TOTAL</b>	82	<b>77</b>		500	<b>520</b>		1,000	<b>1,001</b>		1,500	<b>1,542</b>	
<b>TOTAL18</b>	73	<b>75</b>		386	<b>407</b>		745	<b>743</b>		1,035	<b>1,058</b>	
<b>TOTALHISP</b>	18	<b>11<sup>2</sup></b>	<b>0.0767</b>	18	<b>37<sup>2</sup></b>	<b>0.0352</b>	30	<b>32<sup>3</sup></b>	<b>0.0020</b>	1,237	<b>1,274<sup>1</sup></b>	<b>0.0015</b>
TOTALNH	64	66	0.0767	482	483	0.00352	970	969	0.0020	263	268	0.0015
<b>WHITENH</b>	60	<b>57<sup>1</sup></b>	<b>0.0086</b>	455	<b>462<sup>1</sup></b>	<b>0.0215</b>	306	<b>309<sup>2</sup></b>	<b>0.0027</b>	235	<b>233<sup>2</sup></b>	<b>0.0056</b>
BLACKNH	0	0	0.0000	7	<b>12<sup>3</sup></b>	<b>0.0091</b>	659	<b>650<sup>1</sup></b>	<b>0.0096</b>	10	11	0.0005
AIANNH	4	0	0.0488	6	6	0.0005	4	1	0.0030	0	3	0.0019
ASIANNH	0	<b>2<sup>3</sup></b>	<b>0.0260</b>	11	2	0.0182	0	8	0.0080	18	<b>15<sup>3</sup></b>	<b>0.0023</b>
HPINH	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	2	0.0013
OTHERNH	0	0	0.0000	1	1	0.0000	0	0	0.0000	0	1	0.0006
MLTMNNH	0	7	0.0909	2	0	0.0040	1	1	0.0000	0	3	0.0019
HISP18	14	9	0.0718	10	22	0.0281	21	22	0.0014	807	821	0.0037
NONHISP18	59	66	0.0718	376	385	0.0281	724	721	0.0014	228	237	0.0037
WHITENH18	55	57	0.0066	354	369	0.0105	255	255	0.0000	203	205	0.0024
BLACKNH18	0	0	0.0000	6	7	0.0017	464	461	0.0024	9	10	0.0008
AIANNH18	4	0	0.0548	5	6	0.0018	4	1	0.0040	0	2	0.0019
ASIANNH18	0	2	0.0267	9	2	0.0184	0	4	0.0054	16	15	0.0013
HPINH18	0	0	0.0000	0	0	0.0000	0	0	0.0000	0	2	0.0019
OTHERNH18	0	0	0.0000	0	1	0.0025	0	0	0.0000	0	1	0.0009
MLTMNNH18	0	7	0.0933	2	0	0.0052	1	0	0.0013	0	2	0.0019

Demographic Group (g)	Block Group 040030017032 (AZ)			Block Group 051430110011 (AR)			Block Group 120210112023 (FL)			Block Group 131350505461 (GA)		
	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>
<b>TOTAL</b>	2,000	1,966		3,000	2,939		5,001	5,016		10,000	10,014	
<b>TOTAL18</b>	1,562	1,567		2,153	2,112		3,689	3,697		6,704	6,742	
<b>TOTALHISP</b>	349	336 <sup>2</sup>	0.0036	224	204 <sup>2</sup>	0.0053	1,770	1,806 <sup>2</sup>	0.0061	1,291	1,286 <sup>3</sup>	0.0007
TOTALNH	1,651	1,630	0.0036	2,776	2,735	0.0053	3,231	3,210	0.0061	8,709	8,728	0.0007
<b>WHITENH</b>	1,308	1,324 <sup>1</sup>	0.0194	2,580	2,566 <sup>1</sup>	0.0131	2,891	2,883 <sup>1</sup>	0.0033	3,565	3,571 <sup>2</sup>	0.0001
<b>BLACKNH</b>	181	164 <sup>3</sup>	0.0071	87	73 <sup>3</sup>	0.0042	235	234 <sup>3</sup>	0.0003	4,475	4,482 <sup>1</sup>	0.0001
AIANNH	25	28	0.0017	65	57	0.0023	18	26	0.0016	30	46	0.0016
ASIANNH	106	90	0.0072	32	28	0.0011	59	58	0.0002	473	487	0.0013
HPINH	10	11	0.0006	1	3	0.0007	8	0	0.0016	2	4	0.0002
OTHERNH	3	6	0.0016	4	6	0.0007	7	7	0.0000	79	76	0.0003
MLTMNNH	18	7	0.0054	7	2	0.0017	13	2	0.0022	85	62	0.0023
HISP18	236	233	0.0024	110	96	0.0056	1,193	1,219	0.0063	783	800	0.0019
NONHISP18	1,326	1,334	0.0024	2,043	2,016	0.0056	2,496	2,478	0.0063	5,921	5,942	0.0019
WHITENH18	1,089	1,101	0.0054	1,931	1,920	0.0122	2,267	2,257	0.0040	2,630	2,638	0.0010
BLACKNH18	129	129	0.0003	40	32	0.0034	149	147	0.0006	2,868	2,869	0.0003
AIANNH18	20	24	0.0025	41	40	0.0001	14	21	0.0019	22	34	0.0018
ASIANNH18	72	64	0.0053	23	16	0.0031	50	45	0.0014	304	316	0.0015
HPINH18	4	3	0.0006	1	3	0.0010	4	0	0.0011	2	4	0.0003
OTHERNH18	2	6	0.0025	3	5	0.0010	5	6	0.0003	43	37	0.0009
MLTMNNH18	10	7	0.0019	4	0	0.0019	7	2	0.0014	52	44	0.0012

Demographic Group (g)	Block Group 130510107001 (GA)			Block Group 517100038001 (VA)			Block Group 121199112001 (FL)			Block Group 060730187001 (CA)		
	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>	<i>C<sub>SWA</sub></i>	<i>C<sub>TDA</sub></i>	<i>DR<sub>g</sub></i>
<b>TOTAL</b>	15,089	15,000		19,506	19,517		29,677	29,675		37,452	37,303	
<b>TOTAL18</b>	11,561	11,545		19,486	19,454		29,214	29,198		28,368	28,284	
<b>TOTALHISP</b>	1,066	1,026 <sup>3</sup>	0.0022	2,599	2,581 <sup>3</sup>	0.0010	502	501 <sup>2</sup>	0.0000	8,192	8,091 <sup>2</sup>	0.0018
TOTALNH	14,023	13,974	0.0022	16,907	16,936	0.0010	29,175	29,174	0.0000	29,260	29,212	0.0018
WHITENH	7,901	7,916 <sup>1</sup>	0.0041	10,579	10,599 <sup>1</sup>	0.0007	28,555	28,562 <sup>1</sup>	0.0003	23,326	23,308 <sup>1</sup>	0.0020
BLACKNH	5,281	5,273 <sup>2</sup>	0.0015	4,972	4,975 <sup>2</sup>	0.0000	276	275 <sup>3</sup>	0.0000	3,040	3,040 <sup>3</sup>	0.0003
AIANNH	54	48	0.0004	275	286	0.0006	58	51	0.0002	601	610	0.0003
ASIANNH	643	629	0.0007	776	812	0.0018	246	238	0.0005	1,422	1,420	0.0001
HPINH	17	10	0.0005	80	75	0.0003	7	10	0.0001	340	346	0.0002
OTHERNH	42	32	0.0007	45	39	0.0003	15	10	0.0002	89	74	0.0004
MLTMNNH	85	66	0.0012	180	150	0.0015	18	28	0.0003	442	414	0.0007
HISP18	693	680	0.0010	2,597	2,567	0.0013	460	460	0.0000	5,506	5,449	0.0014
NONHISP18	10,868	10,865	0.0010	16,889	16,887	0.0013	28,754	28,738	0.0000	22,862	22,835	0.0014
WHITENH18	6,404	6,403	0.0007	10,562	10,572	0.0014	28,186	28,193	0.0008	18,751	18,741	0.0016
BLACKNH18	3,849	3,862	0.0016	4,971	4,971	0.0004	247	242	0.0002	2,118	2,107	0.0002
AIANNH18	46	46	0.0000	275	286	0.0006	58	51	0.0002	436	451	0.0006
ASIANNH18	494	486	0.0006	776	799	0.0012	227	213	0.0005	1,032	1,030	0.0000
HPINH18	9	10	0.0001	80	75	0.0003	7	8	0.0000	261	260	0.0000
OTHERNH18	22	19	0.0003	45	37	0.0004	14	10	0.0001	62	54	0.0003
MLTMNNH18	44	39	0.0004	180	147	0.0017	15	21	0.0002	202	192	0.0003

*Motivating Example for Reliable Characteristics*



*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata:

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume
- *TDA* count is **reliable characteristic**

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume
- *TDA* count is **reliable characteristic for the largest demographic group if**

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume
- TDA count is **reliable characteristic for the largest demographic group if its  $DR_g \leq 0.0050$ .**



*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume
- TDA count is **reliable characteristic for the largest demographic group if its  $DR_g \leq 0.0050$ .**

Stratum 1: {0.0086, 0.0215, 0.0096}; No block groups reliable;

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume
- TDA count is **reliable characteristic for the largest demographic group if its  $DR_g \leq 0.0050$ .**

Stratum 1: {0.0086, 0.0215, 0.0096}; No block groups reliable;

Stratum 2: {0.0015, 0.0194, 0.0131 }; 1 out of 3 (0.3333) reliable;

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume
- TDA count is **reliable characteristic for the largest demographic group if its  $DR_g \leq 0.0050$ .**

Stratum 1: {0.0086, 0.0215, 0.0096}; No block groups reliable;

Stratum 2: {0.0015, 0.0194, 0.0131 }; 1 out of 3 (0.3333) reliable;

Stratum 3: {0.0033, 0.0001, 0.0041}; All 3 (1.0000) reliable;

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume
- TDA count is **reliable characteristic for the largest demographic group if its  $DR_g \leq 0.0050$ .**

Stratum 1: {0.0086, 0.0215, 0.0096}; No block groups reliable;

Stratum 2: {0.0015, 0.0194, 0.0131 }; 1 out of 3 (0.3333) reliable;

Stratum 3: {0.0033, 0.0001, 0.0041}; All 3 (1.0000) reliable; and

*Motivating Example for Reliable Characteristics*

- Stratify the 12 block groups we just saw into 4 strata: Show  $DR_g$  for each stratum where  $g$  is largest demographic group and assume
- TDA count is **reliable characteristic for the largest demographic group if its  $DR_g \leq 0.0050$ .**

Stratum 1: {0.0086, 0.0215, 0.0096}; No block groups reliable;

Stratum 2: {0.0015, 0.0194, 0.0131 }; 1 out of 3 (0.3333) reliable;

Stratum 3: {0.0033, 0.0001, 0.0041}; All 3 (1.0000) reliable; and

Stratum 4: {0.0007, 0.0003, 0.0020}. All 3 (1.0000) reliable.

## I.4. THE QUESTION

## I.4. THE QUESTION

What is  $C_{SWA}^*$ ?

## I.4. THE QUESTION

What is  $C_{SWA}^*$ ?

$$C_{SWA(1)} \leq C_{SWA(2)} \leq C_{SWA(3)} \leq \dots \leq C_{SWA}^* \leq \dots \leq C_{SWA(217,739)} \leq C_{SWA(217,740)},$$

(2)



## I.4. THE QUESTION

What is  $C_{SWA}^*$ ?

$$C_{SWA(1)} \leq C_{SWA(2)} \leq C_{SWA(3)} \leq \dots \leq C_{SWA}^* \leq \dots \leq C_{SWA(217,739)} \leq C_{SWA(217,740)},$$

(2)

where the  $C_{SWA(i)}$  counts

## I.4. THE QUESTION

What is  $C_{SWA}^*$ ?

$$C_{SWA(1)} \leq C_{SWA(2)} \leq C_{SWA(3)} \leq \cdots \leq C_{SWA}^* \leq \cdots \leq C_{SWA(217,739)} \leq C_{SWA(217,740)},$$

(2)

where the  $C_{SWA(i)}$  counts are the counts for the TOTAL block group, for  $i = 1; 2; \dots; 217,740$ .

**Table: Proportion of Block Groups in Each Stratum for Three Criteria**  
 (Computations use  $C_{TDA}(g)$  counts that result from 2021-04-28 version of the *TDA*.)  
 Population: United States (50 States & DC)

Stratum for Block Groups Using $C_{SWA}$ for TOTAL	Number of Block Groups	Reliable Characteristics Criteria		
		Criterion I LDG $DR_g \leq 0.01$	Criterion II LDG $DR_g \leq 0.03$	Criterion III LDG $DR_g \leq 0.05$
50 $\leq C_{SWA} < 99$	128	0.1172	0.2812	0.4062
100 $\leq C_{SWA} < 149$	99	0.0909	0.3030	0.4646
150 $\leq C_{SWA} < 199$	124	0.1129	0.3710	0.5565
200 $\leq C_{SWA} < 249$	154	0.2143	0.4545	0.7143
250 $\leq C_{SWA} < 299$	209	0.2105	0.5167	0.7129
300 $\leq C_{SWA} < 349$	264	0.2121	0.5871	0.7803
350 $\leq C_{SWA} < 399$	407	0.2334	0.6757	0.8428
400 $\leq C_{SWA} < 449$	569	0.2900	0.7188	0.8963
450 $\leq C_{SWA} < 499$	915	0.3268	0.7628	0.9355
500 $\leq C_{SWA} < 549$	1,699	0.3431	0.7905	0.9370
<b>550 <math>\leq C_{SWA} &lt; 599</math></b>	<b>3,238</b>	0.3811	0.8295	<b>0.9580</b>
600 $\leq C_{SWA} < 649$	5,131	0.3962	0.8564	0.9723
650 $\leq C_{SWA} < 699$	6,683	0.4200	0.8692	0.9753
700 $\leq C_{SWA} < 749$	7,356	0.4468	0.8802	0.9826
750 $\leq C_{SWA} < 799$	8,170	0.4477	0.8973	0.9838
800 $\leq C_{SWA} < 849$	8,213	0.4785	0.9190	0.9907
850 $\leq C_{SWA} < 899$	8,441	0.4971	0.9231	0.9892
900 $\leq C_{SWA} < 949$	8,657	0.5021	0.9287	0.9928
950 $\leq C_{SWA} < 999$	8,723	0.5202	0.9411	0.9948
1,000 $\leq C_{SWA} < 1,049$	8,398	0.5460	0.9447	0.9936
<b>1,050 <math>\leq C_{SWA} &lt; 1,099</math></b>	<b>8,345</b>	0.5464	<b>0.9575</b>	0.9959
1,100 $\leq C_{SWA} < 1,149$	7,950	0.5552	0.9572	0.9969
1,150 $\leq C_{SWA} < 1,199$	7,860	0.5748	0.9626	0.9971

Table (Continued):

				Reliable Characteristics Criteria		
Stratum for Block Groups Using $C_{SWA}$ for TOTAL	Number of Block Groups			Criterion I	Criterion II	Criterion III
				$LDG DR_g \leq 0.01$	$LDG DR_g \leq 0.03$	$LDG DR_g \leq 0.05$
1,200	$C_{SWA}$	1,249	7,451	0.5770	0.9691	0.9977
1,250	$C_{SWA}$	1,299	7,124	0.6049	0.9698	0.9983
1,300	$C_{SWA}$	1,349	6,714	0.6151	0.9724	0.9993
1,350	$C_{SWA}$	1,399	6,507	0.6178	0.9743	0.9989
1,400	$C_{SWA}$	1,449	5,911	0.6287	0.9785	0.9980
1,450	$C_{SWA}$	1,499	5,617	0.6386	0.9810	0.9993
1,500	$C_{SWA}$	1,549	5,390	0.6471	0.9848	0.9996
1,550	$C_{SWA}$	1,599	4,856	0.6623	0.9841	0.9992
1,600	$C_{SWA}$	1,649	4,508	0.6528	0.9878	0.9998
1,650	$C_{SWA}$	1,699	4,325	0.6805	0.9864	0.9998
1,700	$C_{SWA}$	1,749	4,093	0.6895	0.9924	0.9993
1,750	$C_{SWA}$	1,799	3,689	0.6837	0.9883	0.9997
1,800	$C_{SWA}$	1,849	3,469	0.7094	0.9928	0.9997
1,850	$C_{SWA}$	1,899	3,252	0.7011	0.9889	1.0000
1,900	$C_{SWA}$	1,949	3,008	0.7048	0.9924	0.9997
1,950	$C_{SWA}$	1,999	2,832	0.7334	0.9926	0.9996
2,000	$C_{SWA}$	2,049	2,573	0.7178	0.9953	1.0000
2,050	$C_{SWA}$	2,099	2,356	0.7394	0.9949	1.0000
2,100	$C_{SWA}$	2,149	2,307	0.7391	0.9944	0.9991
2,150	$C_{SWA}$	2,199	2,033	0.7634	0.9970	1.0000
2,200	$C_{SWA}$	2,249	1,999	0.7564	0.9970	0.9995
2,250	$C_{SWA}$	2,299	1,892	0.7627	0.9963	1.0000
2,300	$C_{SWA}$	2,349	1,666	0.7533	0.9976	0.9994
2,350	$C_{SWA}$	2,399	1,622	0.7608	0.9957	1.0000
2,400	$C_{SWA}$	2,449	1,421	0.7643	0.9986	1.0000
2,450	$C_{SWA}$	2,499	1,350	0.7733	0.9970	0.9993
Total			199,698			

Using public released data

Using public released data (one run of the 2021-04-28 version of *TDA*),

Using public released data (one run of the 2021-04-28 version of *TDA*), we might say,

Using public released data (one run of the 2021-04-28 version of *TDA*), we might say, empirically based on the data for the block groups used in our study,



Using public released data (one run of the 2021-04-28 version of *TDA*), we might say, empirically based on the data for the block groups used in our study, that

*“for any block group*

Using public released data (one run of the 2021-04-28 version of *TDA*), we might say, empirically based on the data for the block groups used in our study, that

*“for any block group with a TOTAL count near 600 people,*

Using public released data (one run of the 2021-04-28 version of *TDA*), we might say, empirically based on the data for the block groups used in our study, that

*“for any block group with a TOTAL count near 600 people, the difference between the TDA ratio of the largest demographic group (LDG)*

Using public released data (one run of the 2021-04-28 version of TDA), we might say, empirically based on the data for the block groups used in our study, that

*“for any block group with a TOTAL count near 600 people, the difference between the TDA ratio of the largest demographic group (LDG) and the corresponding SWA ratio for the LDG*

Using public released data (one run of the 2021-04-28 version of TDA), we might say, empirically based on the data for the block groups used in our study, that

*“for any block group with a TOTAL count near 600 people, the difference between the TDA ratio of the largest demographic group (LDG) and the corresponding SWA ratio for the LDG is less than or equal to 5 percentage points*

Using public released data (one run of the 2021-04-28 version of TDA), we might say, empirically based on the data for the block groups used in our study, that

*“for any block group with a TOTAL count near 600 people, the difference between the TDA ratio of the largest demographic group (LDG) and the corresponding SWA ratio for the LDG is less than or equal to 5 percentage points at least 95% of the time”.*

Applied same version of *TDA*

Applied same version of *TDA* 25 independent times (runs) to CEF.



Applied same version of *TDA* 25 independent times (runs) to CEF.  
 Stratum for each run, where 0.9500 was exceeded is in Table.

**Table:** For Each Run, the Stratum and Stratum Proportion When 0.9500 First Exceeded  
 Population: United States (50 States & DC)

<i>TDA</i> Run	Stratum for Block Groups			Criterion III LDG $DR_g < 0.05$ Proportion When 0.9500 First Exceeded
	1	550	C <sub>SWA</sub>	599
2	550	C <sub>SWA</sub>	599	0.9605
3	550	C <sub>SWA</sub>	599	0.9623
4	550	C <sub>SWA</sub>	599	0.9642
5	550	C <sub>SWA</sub>	599	0.9608
6	550	C <sub>SWA</sub>	599	0.9580
7	550	C <sub>SWA</sub>	599	0.9592
8	550	C <sub>SWA</sub>	599	0.9614
9	550	C <sub>SWA</sub>	599	0.9595
10	550	C <sub>SWA</sub>	599	0.9636
11	550	C <sub>SWA</sub>	599	0.9592
12	550	C <sub>SWA</sub>	599	0.9589
13	550	C <sub>SWA</sub>	599	0.9592
14	550	C <sub>SWA</sub>	599	0.9617
15	550	C <sub>SWA</sub>	599	0.9589
16	550	C <sub>SWA</sub>	599	0.9617
17	550	C <sub>SWA</sub>	599	0.9617
18	550	C <sub>SWA</sub>	599	0.9614
19	550	C <sub>SWA</sub>	599	0.9592
20	550	C <sub>SWA</sub>	599	0.9558
21	550	C <sub>SWA</sub>	599	0.9592
22	550	C <sub>SWA</sub>	599	0.9589
23	550	C <sub>SWA</sub>	599	0.9580
24	550	C <sub>SWA</sub>	599	0.9611
25	550	C <sub>SWA</sub>	599	0.9568

"Place and MCD" (21,00+ entities) as Alternative to "Block Group"

**Table:** For Each Run, the Stratum and Stratum Proportion When 0.9500 First Exceeded  
Population: United States (50 States & DC)

TDA Run	Stratum for Places & MCDs	Criterion III
		LDG $DR_g \leq 0.05$ Proportion When 0.9500 First Exceeded
1	300 \ C <sub>S</sub> WA \ 349	0.9621
2	250 \ C <sub>S</sub> WA \ 299	0.9580
3	300 \ C <sub>S</sub> WA \ 349	0.9598
4	250 \ C <sub>S</sub> WA \ 299	0.9580
5	300 \ C <sub>S</sub> WA \ 349	0.9665
6	300 \ C <sub>S</sub> WA \ 349	0.9688
7	300 \ C <sub>S</sub> WA \ 349	0.9688
8	300 \ C <sub>S</sub> WA \ 349	0.9621
9	300 \ C <sub>S</sub> WA \ 349	0.9754
10	300 \ C <sub>S</sub> WA \ 349	0.9576
11	300 \ C <sub>S</sub> WA \ 349	0.9598
12	300 \ C <sub>S</sub> WA \ 349	0.9777
13	300 \ C <sub>S</sub> WA \ 349	0.9598
14	300 \ C <sub>S</sub> WA \ 349	0.9688
15	300 \ C <sub>S</sub> WA \ 349	0.9688
16	300 \ C <sub>S</sub> WA \ 349	0.9643
17	300 \ C <sub>S</sub> WA \ 349	0.9732
18	300 \ C <sub>S</sub> WA \ 349	0.9665
19	300 \ C <sub>S</sub> WA \ 349	0.9710
20	300 \ C <sub>S</sub> WA \ 349	0.9621
21	300 \ C <sub>S</sub> WA \ 349	0.9688
22	350 \ C <sub>S</sub> WA \ 399	0.9520
23	300 \ C <sub>S</sub> WA \ 349	0.9643
24	300 \ C <sub>S</sub> WA \ 349	0.9598
25	300 \ C <sub>S</sub> WA \ 349	0.9732

*"Congressional & State Legislative District" as Alternative to "Block Group"*

- “Congressional & State Legislative District” as Alternative to “Block Group”*
- Congressional district(s) (CD)

*“Congressional & State Legislative District” as Alternative to “Block Group”*

- Congressional district(s) (CD)
- State legislative districts in an upper chamber (SLDU)

*“Congressional & State Legislative District” as Alternative to “Block Group”*

- Congressional district(s) (CD)
- State legislative districts in an upper chamber (SLDU)
- State legislative districts in a lower chamber (SLDL)

*“Congressional & State Legislative District” as Alternative to “Block Group”*

- Congressional district(s) (CD)
- State legislative districts in an upper chamber (SLDU)
- State legislative districts in a lower chamber (SLDL)

	CD	SLDU	SLDL
Number of Districts in U.S.	436	1,946	4,785
Min Population	526,283	13,629	3,173
Median Population	705,831	121,212	41,713
Mean Population	708,132	158,656	64,016
Max Population	989,415	940,612	470,325

**Table:** For Each Run, the Stratum and Stratum Proportion When 0.9500 First Exceeded  
Population: United States (50 States & DC)

TDA Run	Stratum for		Criterion III	
	Congressional & State	Legislative Districts	LDG $DR_g \leq 0.05$ Proportion When 0.9500 First Exceeded	
1	3,150	C <sub>SWA</sub>	3,199	1.0000
2	3,150	C <sub>SWA</sub>	3,199	1.0000
3	3,150	C <sub>SWA</sub>	3,199	1.0000
4	3,150	C <sub>SWA</sub>	3,199	1.0000
5	3,150	C <sub>SWA</sub>	3,199	1.0000
6	3,150	C <sub>SWA</sub>	3,199	1.0000
7	3,150	C <sub>SWA</sub>	3,199	1.0000
8	3,150	C <sub>SWA</sub>	3,199	1.0000
9	3,150	C <sub>SWA</sub>	3,199	1.0000
10	3,150	C <sub>SWA</sub>	3,199	1.0000
11	3,150	C <sub>SWA</sub>	3,199	1.0000
12	3,150	C <sub>SWA</sub>	3,199	1.0000
13	3,150	C <sub>SWA</sub>	3,199	1.0000
14	3,150	C <sub>SWA</sub>	3,199	1.0000
15	3,150	C <sub>SWA</sub>	3,199	1.0000
16	3,150	C <sub>SWA</sub>	3,199	1.0000
17	3,150	C <sub>SWA</sub>	3,199	1.0000
18	3,150	C <sub>SWA</sub>	3,199	1.0000
19	3,150	C <sub>SWA</sub>	3,199	1.0000
20	3,150	C <sub>SWA</sub>	3,199	1.0000
21	3,150	C <sub>SWA</sub>	3,199	1.0000
22	3,150	C <sub>SWA</sub>	3,199	1.0000
23	3,150	C <sub>SWA</sub>	3,199	1.0000
24	3,150	C <sub>SWA</sub>	3,199	1.0000
25	3,150	C <sub>SWA</sub>	3,199	1.0000



## I.5. CONCLUDING REMARKS FOR PART I

## I.5. CONCLUDING REMARKS FOR PART I

Remark 1:

## I.5. CONCLUDING REMARKS FOR PART I

### Remark 1:

- $C_{SWA}^*$  is an empirical result.

## I.5. CONCLUDING REMARKS FOR PART I

### Remark 1:

- $C_{SWA}^*$  is an empirical result.
- Seems to hold for (1) block groups; (2) places and MCDs; (3) congressional and state legislative districts.

## I.5. CONCLUDING REMARKS FOR PART I

### Remark 1:

- $C_{SWA}^*$  is an empirical result.
- Seems to hold for (1) block groups; (2) places and MCDs; (3) congressional and state legislative districts.

### Remark 2:

## I.5. CONCLUDING REMARKS FOR PART I

### **Remark 1:**

- $C_{SWA}^*$  is an empirical result.
- Seems to hold for (1) block groups; (2) places and MCDs; (3) congressional and state legislative districts.

### **Remark 2:**

- While small demographic groups are important,

## I.5. CONCLUDING REMARKS FOR PART I

### **Remark 1:**

- $C_{SWA}^*$  is an empirical result.
- Seems to hold for (1) block groups; (2) places and MCDs; (3) congressional and state legislative districts.

### **Remark 2:**

- While small demographic groups are important, in the context of redistricting,

## I.5. CONCLUDING REMARKS FOR PART I

### **Remark 1:**

- $C_{SWA}^*$  is an empirical result.
- Seems to hold for (1) block groups; (2) places and MCDs; (3) congressional and state legislative districts.

### **Remark 2:**

- While small demographic groups are important, in the context of redistricting, it is the largest among the demographic groups that have the potential to form districts



## I.5. CONCLUDING REMARKS FOR PART I

### **Remark 1:**

- $C_{SWA}^*$  is an empirical result.
- Seems to hold for (1) block groups; (2) places and MCDs; (3) congressional and state legislative districts.

### **Remark 2:**

- While small demographic groups are important, in the context of redistricting, it is the largest among the demographic groups that have the potential to form districts where sufficiently large (and compact) minority groups

## I.5. CONCLUDING REMARKS FOR PART I

### **Remark 1:**

- $C_{SWA}^*$  is an empirical result.
- Seems to hold for (1) block groups; (2) places and MCDs; (3) congressional and state legislative districts.

### **Remark 2:**

- While small demographic groups are important, in the context of redistricting, it is the largest among the demographic groups that have the potential to form districts where sufficiently large (and compact) minority groups have the opportunity

## I.5. CONCLUDING REMARKS FOR PART I

### **Remark 1:**

- $C_{SWA}^*$  is an empirical result.
- Seems to hold for (1) block groups; (2) places and MCDs; (3) congressional and state legislative districts.

### **Remark 2:**

- While small demographic groups are important, in the context of redistricting, it is the largest among the demographic groups that have the potential to form districts where sufficiently large (and compact) minority groups have the opportunity “to elect representatives of their choice”.

## Part II

## Part II

### II.1. INTRODUCTION

## Part II

### II.1. INTRODUCTION

Part II.

- Update of earlier study in [5] where  $\epsilon = 4.0$

## Part II

### II.1. INTRODUCTION

Part II.

- Update of earlier study in [5] where  $\epsilon = 4.0$  and the 2019-10-31 version of *TDA* was used;

## Part II

### II.1. INTRODUCTION

Part II.

- Update of earlier study in [5] where  $\epsilon = 4.0$  and the 2019-10-31 version of *TDA* was used;
- In this study,



## Part II

### II.1. INTRODUCTION

Part II.

- Update of earlier study in [5] where  $\epsilon = 4.0$  and the 2019-10-31 version of *TDA* was used;
- In this study,  $\epsilon = 10.3$

## Part II

### II.1. INTRODUCTION

Part II.

- Update of earlier study in [5] where  $\epsilon = 4.0$  and the 2019-10-31 version of *TDA* was used;
- In this study,  $\epsilon = 10.3$  and advances have been made resulting in the 2021-04-28 version of *TDA*.

## 2010 Census Data for Rhode Island

## 2010 Census Data for Rhode Island

Rhode Island has:

## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),

## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),
- 38 state legislative districts (SLDU) in its upper legislative chamber, and

## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),
- 38 state legislative districts (SLDU) in its upper legislative chamber, and
- 75 state legislative districts (SLDL) in its lower legislative chamber.

## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),
- 38 state legislative districts (SLDU) in its upper legislative chamber, and
- 75 state legislative districts (SLDL) in its lower legislative chamber.

## **2010 Census Data for Three Cases Provided by DOJ**



## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),
- 38 state legislative districts (SLDU) in its upper legislative chamber, and
- 75 state legislative districts (SLDL) in its lower legislative chamber.

## **2010 Census Data for Three Cases Provided by DOJ**

Conduct similar analyses of data in

## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),
- 38 state legislative districts (SLDU) in its upper legislative chamber, and
- 75 state legislative districts (SLDL) in its lower legislative chamber.

## **2010 Census Data for Three Cases Provided by DOJ**

Conduct similar analyses of data in

- Panola County, Mississippi (MS) (2,180 blocks);

## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),
- 38 state legislative districts (SLDU) in its upper legislative chamber, and
- 75 state legislative districts (SLDL) in its lower legislative chamber.

## **2010 Census Data for Three Cases Provided by DOJ**

Conduct similar analyses of data in

- Panola County, Mississippi (MS) (2,180 blocks);
- Tate County (School District), MS (784 blocks);

## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),
- 38 state legislative districts (SLDU) in its upper legislative chamber, and
- 75 state legislative districts (SLDL) in its lower legislative chamber.

## **2010 Census Data for Three Cases Provided by DOJ**

Conduct similar analyses of data in

- Panola County, Mississippi (MS) (2,180 blocks);
- Tate County (School District), MS (784 blocks); and

## **2010 Census Data for Rhode Island**

Rhode Island has:

- two (2) congressional districts (CD),
- 38 state legislative districts (SLDU) in its upper legislative chamber, and
- 75 state legislative districts (SLDL) in its lower legislative chamber.

## **2010 Census Data for Three Cases Provided by DOJ**

Conduct similar analyses of data in

- Panola County, Mississippi (MS) (2,180 blocks);
- Tate County (School District), MS (784 blocks); and
- Tylertown (Walthall County), MS (136 blocks).

		2010 Census, SF1 (PE 91-172)(2013)				Counts & Percentages, 113 <sup>th</sup> Congress 3 Out of 25 Runs of the TDA					
Demographics		113 <sup>th</sup> Congress		TDA-Run A		TDA-Run B		TDA-Run C			
DIST-ID	Rhode Island	CD-01	CD-02	CD-01	CD-02	CD-01	CD-02	CD-01	CD-02		
TOTAL	1,052,567	526,283	526,284	526,449	526,118	526,173	526,394	525,872	526,695		
DEV		-0.5	0.5	165.5	-165.5	-110.5	110.5	-411.5	411.5		
DEVP		0.00	0.00	0.03	-0.03	-0.02	0.02	-0.08	0.08		
TOTAL18	828,611	412,778	415,833	412,736	415,826	412,776	415,807	412,512	416,054		
TOTALHISP	130,655	76,100	54,555	76,248	54,402	76,230	54,402	76,153	54,539		
TOTALHISPP	12.41	14.46	10.37	14.48	10.34	14.49	10.33	14.48	10.35		
TOTALNH	921,912	450,183	471,729	450,201	471,716	449,943	471,992	449,719	472,156		
TOTALNHP	87.59	85.54	89.63	85.52	89.66	85.51	89.67	85.52	89.65		
WHITENHP	803,685	377,109	426,576	377,022	426,658	376,955	426,735	377,012	426,677		
WHITENHP	76.35	71.66	81.05	71.62	81.10	71.64	81.07	71.69	81.01		
BLACKNH	57,927	37,627	20,300	37,704	20,219	37,705	20,247	37,517	20,406		
BLACKNHP	5.50	7.15	3.86	7.16	3.84	7.17	3.85	7.13	3.87		
AIANNH	6,839	3,142	3,697	3,201	3,672	3,126	3,717	3,141	3,735		
AIANNHP	0.65	0.60	0.70	0.61	0.70	0.59	0.71	0.60	0.71		
ASIANNH	34,194	17,705	16,489	17,692	16,505	17,684	16,496	17,723	16,478		
ASIANNHP	3.25	3.36	3.13	3.36	3.14	3.36	3.13	3.37	3.13		
HPINH	655	383	272	427	242	400	263	355	293		
HPINH	0.06	0.07	0.05	0.08	0.05	0.08	0.05	0.07	0.06		
OTHERNH	10,296	8,492	1,804	8,443	1,845	8,454	1,845	8,457	1,829		
OTHERNHP	0.98	1.61	0.34	1.60	0.35	1.61	0.35	1.61	0.35		
MLTMNNH	8,316	5,725	2,591	5,712	2,575	5,619	2,689	5,514	2,738		
MLTMNHP	0.79	1.09	0.49	1.09	0.49	1.07	0.51	1.05	0.52		
HISP18	84,715	49,303	35,412	49,333	35,349	49,428	35,253	49,331	35,368		
HISP18P	10.22	11.94	8.52	11.95	8.50	11.97	8.48	11.96	8.50		
NONHISP18	743,896	363,475	380,421	363,403	380,477	363,348	380,554	363,181	380,686		
NONHISP18P	89.78	88.06	91.48	88.05	91.50	88.03	91.52	88.04	91.50		
WHITENH18	690,823	312,240	348,583	312,178	348,640	312,163	348,684	312,232	348,589		
WHITENH18P	79.75	75.64	83.83	75.64	83.84	75.63	83.86	75.69	83.78		
BLACKNH18	39,485	25,402	14,083	25,414	14,060	25,425	14,068	25,326	14,153		
BLACKNH18P	4.77	6.15	3.39	6.16	3.38	6.16	3.38	6.14	3.40		
AIANNH18	4,963	2,332	2,631	2,326	2,645	2,291	2,666	2,317	2,670		
AIANNH18P	0.60	0.56	0.63	0.56	0.64	0.56	0.64	0.56	0.64		
ASIANNH18	25,333	13,276	12,057	13,229	12,106	13,282	12,035	13,326	12,008		
ASIANNH18P	3.06	3.22	2.90	3.21	2.91	3.22	2.89	3.23	2.89		
HPINH18	500	307	193	334	175	313	195	275	221		
HPINH18P	0.06	0.07	0.05	0.08	0.04	0.08	0.05	0.07	0.05		
OTHERNH18	7,290	6,061	1,229	6,059	1,224	6,067	1,214	6,008	1,271		
OTHERNH18P	0.88	1.47	0.30	1.47	0.29	1.47	0.29	1.46	0.31		
MLTMNH18	5,502	3,857	1,645	3,863	1,627	3,807	1,692	3,697	1,774		
MLTMNH18P	0.66	0.93	0.40	0.94	0.39	0.92	0.41	0.90	0.43		

Source: Data from 3 Runs of the TDA, U. S. Bureau of the Census, Washington, D.C.

Demographics	2010 Census, SF1 (PL 94-171) (2013) Counts & Percentages POST-2010 Plan				Counts & Percentages, 2013 Run A of the TDA			
	SLDU-01	SLDU-02	SLDU-03	SLDU-04	SLDU-01	SLDU-02	SLDU-03	SLDU-04
TOTAL	28,161	28,079	28,398	28,201	27,836	27,823	28,716	28,201
DEV	461.9	379.9	698.9	501.9	136.9	123.9	1,016.9	501.9
DEVP	1.64	1.35	2.46	1.78	0.49	0.45	3.54	1.78
TOTAL18	20,914	19,846	25,361	23,599	20,746	19,706	25,506	23,592
TOTALHISP	10,282	16,288	1,409	3,217	10,142	16,134	1,525	3,192
TOTALHISPSP	36.51	58.01	4.96	11.41	36.43	57.99	5.31	11.32
TOTALNH	17,879	11,791	26,989	24,984	17,694	11,689	27,191	25,009
TOTALNHP	63.49	41.99	95.04	88.59	63.57	42.01	94.69	88.68
WHITENH	10,222	3,553	22,028	21,210	10,216	3,531	22,030	21,305
WHITENHP	36.30	12.65	77.57	75.21	36.70	12.69	76.72	75.55
BLACKNH	4,862	4,332	1,124	2,348	4,814	4,309	1,164	2,318
BLACKNHP	17.27	15.43	3.96	8.33	17.29	15.49	4.05	8.22
AIANNH	283	216	135	172	254	186	170	170
AIANNHP	1.00	0.77	0.48	0.61	0.91	0.67	0.59	0.60
ASIANNH	1,526	3,032	3,262	826	1,587	3,051	5,253	781
ASIANNHP	5.42	10.80	11.49	2.93	5.70	10.97	11.33	2.77
HPINH	25	11	16	14	18	6	27	9
HPINHHP	0.09	0.04	0.06	0.05	0.06	0.02	0.09	0.03
OTHERNH	457	189	224	241	438	196	253	220
OTHERNHP	1.62	0.67	0.79	0.85	1.57	0.70	0.88	0.78
MLTMNH	594	458	200	173	367	410	291	206
MLTMNHP	1.79	1.63	0.70	0.61	1.32	1.47	1.02	0.73
HISP18	6,458	11,014	1,241	2,097	6,369	10,919	1,262	2,088
HISP18P	30.88	55.50	4.89	8.89	30.70	55.41	4.95	8.85
NONHISP18	14,456	8,832	24,120	21,502	14,377	8,787	24,244	21,504
NONHISP18P	69.12	44.50	95.11	91.11	69.30	44.59	95.05	91.15
WHITENH18	9,431	3,062	19,682	18,839	9,134	3,049	19,703	18,919
WHITENH18P	43.66	15.43	77.61	79.83	44.03	15.47	77.25	80.19
BLACKNH18	3,309	3,027	973	1,599	3,279	3,006	990	1,585
BLACKNH18P	15.82	15.25	3.84	6.78	15.81	15.25	3.88	6.72
AIANNH18	197	154	110	136	186	140	123	123
AIANNH18P	0.94	0.78	0.43	0.58	0.90	0.71	0.48	0.52
ASIANNH18	1,170	2,135	2,989	611	1,197	2,160	2,980	577
ASIANNH18P	5.59	10.76	11.79	2.59	5.77	10.96	11.68	2.45
HPINH18	20	11	14	13	11	5	21	5
HPINH18P	0.10	0.06	0.06	0.06	0.05	0.03	0.08	0.02
OTHERNH18	326	125	186	178	325	125	201	170
OTHERNH18P	1.56	0.63	0.73	0.75	1.57	0.63	0.79	0.72
MLTMNH18	303	318	166	126	245	302	226	125
MLTMNH18P	1.45	1.60	0.65	0.53	1.18	1.53	0.89	0.53

Demographics	2010 Census, SF1 (PL 94-171) (2013) Counts & Percentages POST-2010 Plan				Counts & Percentages, 2013 Run A of the TDA			
	SLDL-01	SLDL-02	SLDL-03	SLDL-04	SLDL-01	SLDL-02	SLDL-03	SLDL-04
TOTAL	13,881	13,821	13,949	13,713	14,072	13,707	13,714	13,660
DEV	-153.2	-213.2	-85.2	-321.2	37.8	-327.2	-320.2	-374.2
DEVP	-1.10	-1.54	-0.61	-2.34	0.27	-2.39	-2.34	-2.74
TOTAL18	12,835	12,800	9,607	11,205	12,899	12,699	9,523	11,166
TOTALHISP	1,902	1,768	5,905	1,049	1,086	1,692	5,826	1,033
TOTALHISPSP	7.22	12.79	42.33	7.65	7.72	12.34	42.48	7.56
TOTALNH	12,879	12,053	8,044	12,664	12,986	12,015	7,888	12,627
TOTALNHP	92.78	87.21	57.67	92.35	92.28	87.66	57.52	92.44
WHITENH	9,922	8,714	3,465	9,539	9,899	8,697	3,464	9,547
WHITENHP	71.48	63.05	24.84	69.56	70.35	63.45	25.26	69.89
BLACKNH	581	1,125	3,015	1,495	605	1,128	2,969	1,509
BLACKNHP	4.19	8.14	21.61	10.90	4.30	8.23	21.65	11.05
AIANNH	46	104	189	126	66	123	152	99
AIANNHP	0.33	0.75	1.35	0.92	0.47	0.90	1.11	0.72
ASIANNH	2,175	1,776	794	792	2,167	1,753	823	803
ASIANNHP	15.67	12.85	5.69	5.78	15.40	12.79	6.00	5.88
HPINH	12	16	12	1	25	11	6	9
HPINHP	0.09	0.12	0.09	0.01	0.18	0.08	0.04	0.07
OTHERNH	57	148	257	396	85	130	240	392
OTHERNHP	0.41	1.07	1.84	2.89	0.60	0.95	1.75	2.87
MLTMNNH	86	170	312	315	139	173	234	268
MLTMNNHP	0.62	1.23	2.24	2.30	0.99	1.26	1.71	1.96
HISP18	951	1,475	3,518	693	977	1,398	3,498	666
HISP18P	7.41	11.52	36.62	6.18	7.57	11.01	36.73	5.96
NONHISP18	11,884	11,325	6,089	10,512	11,922	11,301	6,025	10,500
NONHISP18P	92.59	88.48	63.38	93.82	92.43	88.59	63.27	94.04
WHITENH18	9,981	8,209	3,040	8,119	9,968	8,338	3,038	8,137
WHITENH18P	70.75	65.15	31.64	72.46	70.30	65.66	31.90	72.87
BLACKNH18	569	972	1,971	1,144	557	976	1,945	1,163
BLACKNH18P	4.36	7.59	20.52	10.21	4.32	7.69	20.42	10.42
AIANNH18	45	82	129	101	50	99	110	85
AIANNH18P	0.35	0.64	1.34	0.90	0.39	0.78	1.16	0.76
ASIANNH18	2,052	1,655	575	635	2,037	1,633	589	644
ASIANNH18P	15.99	12.93	5.99	5.67	15.79	12.86	6.19	5.77
HPINH18	10	14	11	1	22	8	2	3
HPINH18P	0.08	0.11	0.11	0.01	0.17	0.06	0.02	0.03
OTHERNH18	51	126	190	280	69	110	181	281
OTHERNH18P	0.40	0.98	1.98	2.50	0.53	0.87	1.90	2.52
MLTMNH18	85	137	173	232	119	137	160	187
MLTMNH18P	0.66	1.07	1.80	2.07	0.92	1.08	1.68	1.67



$$2010 \text{ Census IDEAL POPULATION} = \frac{34,707}{5} = 6,941.4$$

$$TDA \text{ IDEAL POPULATION} = \frac{34,702}{5} = 6,940.4$$

Demographics	2010 Census, SF1 (PL 94-171) Counts & Percentages POST-2010 Plan										Counts & Percentages Run A of the TDA											
	Panels	01					02					Panels	01					02				
		01	02	03	04	05	01	02	03	04	05		01	02	03	04	05					
TOTAL	34,707	6,974	6,549	7,074	7,105	7,005	34,702	7,044	6,571	7,033	7,066	6,968										
DEV		32.6	-392.4	132.6	163.6	63.6		103.6	-369.4	92.6	125.6	47.6										
DEVP		0.47	-5.99	1.87	2.30	0.91		1.47	-5.62	1.32	1.78	0.68										
TOTALIS	25,363	5,214	4,732	5,171	5,345	4,901	25,384	5,267	4,730	5,171	5,313	4,903										
TOTALHISP		494	66	75	85	120	148		521	98	80	104	159									
TOTALHISPSP		1.42	0.95	1.15	1.20	1.69	2.11		1.50	1.39	1.22	1.14	1.47	2.28								
TOTALNH		34,213	6,908	6,474	6,989	6,985	6,857		34,181	6,946	6,491	6,953	6,962	6,829								
TOTALNHSP		98.58	99.05	98.85	98.80	98.31	97.89		98.50	98.61	98.78	98.86	98.53	97.72								
WHITENH		16,981	2,419	2,096	4,030	5,250	3,186		16,989	2,455	2,084	4,029	5,249	3,181								
WHITENHSP		48.93	34.69	32.00	56.97	73.89	45.48		48.96	34.85	31.72	57.16	74.29	45.52								
BLACKNH		16,899	4,427	4,332	2,925	1,658	3,557		16,870	4,421	4,345	2,890	1,690	3,551								
BLACKNHSP		48.69	63.48	66.15	41.35	23.34	59.78		48.61	62.76	66.12	41.13	23.49	59.82								
AIANNH		148	26	20	15	38	49		143	28	24	21	34	36								
AIANNHSP		0.43	0.37	0.31	0.21	0.53	0.70		0.41	0.40	0.37	0.30	0.48	0.52								
ASIANNH		89	8	7	5	17	52		100	14	20	8	9	49								
ASIANNHSP		0.26	0.11	0.11	0.07	0.24	0.74		0.29	0.20	0.30	0.11	0.13	0.70								
HPNH		4	0	0	0	2	2		0	0	0	0	0	0								
HPNHSP		0.01	0.00	0.00	0.00	0.03	0.03		0.00	0.00	0.00	0.00	0.00	0.00								
OTHERNH		19	7	5	1	3	3		4	2	2	0	0	0								
OTHERNHSP		0.05	0.10	0.08	0.01	0.04	0.04		0.01	0.03	0.03	0.00	0.00	0.00								
MLTMNNH		73	21	14	13	17	8		75	26	16	11	10	12								
MLTMNNHSP		0.21	0.30	0.21	0.18	0.24	0.11		0.22	0.37	0.24	0.16	0.14	0.17								
HISP18		298	44	44	52	63	95		230	71	57	43	61	88								
HISP18SP		1.17	0.84	0.93	1.01	1.18	1.94		1.26	1.35	1.21	0.83	1.15	1.79								
NONHISP18		25,065	5,170	4,688	5,119	5,382	4,806		25,064	5,196	4,673	5,128	5,252	4,815								
NONHISP18SP		98.83	99.16	99.07	98.99	98.82	98.06		98.74	98.65	98.79	99.17	98.85	98.21								
WHITENH18		13,455	2,025	1,732	3,072	4,115	2,511		13,464	2,044	1,697	3,097	4,112	2,514								
WHITENH18SP		53.05	38.84	36.60	59.41	70.99	51.23		53.04	38.81	35.88	59.89	77.40	51.27								
BLACKNH18		11,394	3,099	2,928	2,024	1,118	2,225		11,386	3,110	2,937	2,004	1,107	2,228								
BLACKNH18SP		44.92	59.44	61.88	39.14	20.92	45.40		44.86	59.05	62.09	38.75	20.84	45.44								
AIANNH18		115	21	16	11	29	38		116	22	18	17	23	36								
AIANNH18SP		0.45	0.40	0.34	0.21	0.54	0.78		0.46	0.42	0.38	0.33	0.43	0.73								
ASIANNH18		54	8	5	2	12	27		60	7	13	4	4	32								
ASIANNH18SP		0.21	0.15	0.11	0.04	0.22	0.55		0.24	0.13	0.27	0.08	0.08	0.65								
HPNH18		2	0	0	0	1	1		0	0	0	0	0	0								
HPNH18SP		0.01	0.00	0.00	0.00	0.02	0.02		0.00	0.00	0.00	0.00	0.00	0.00								
OTHERNH18		5	1	0	1	2	1		0	0	0	0	0	0								
OTHERNH18SP		0.02	0.02	0.00	0.02	0.04	0.02		0.00	0.00	0.00	0.00	0.00	0.00								
MLTMNH18		40	16	7	9	5	3		38	13	8	6	6	5								
MLTMNH18SP		0.16	0.31	0.15	0.17	0.09	0.06		0.15	0.25	0.17	0.12	0.11	0.10								

$$2010 \text{ Census IDEAL POPULATION} = \frac{18,823}{5} = 3,764.6$$

$$TDA \text{ IDEAL POPULATION} = \frac{18,831}{5} = 3,766.2$$

Demographics	2010 Census, SFI (PL 94-171)										Counts & Percentages Run A of the TDA									
	Counts & Percentages POST-2010 Plan										Counts & Percentages POST-2010 Plan									
	Year	01	02	03	04	05	Year	01	02	03	04	05	Year	01	02	03	04	05		
TOTAL	18,823	3,014	3,803	3,626	3,697	3,684	18,831	3,019	3,886	3,654	3,750	3,622								
DEV		149.4	128.4	-99.6	-67.6	-110.6		152.8	119.8	-112.2	-16.2	-144.2								
DEVP		3.82	3.30	-2.72	-1.83	-3.03		3.90	3.08	-3.07	-0.43	-3.98								
TOTALIS	13,893	2,780	2,826	2,799	2,755	2,733	13,900	2,788	2,833	2,796	2,773	2,719								
TOTALHSP	399	87	63	110	32	107	388	87	70	102	57	72								
TOTALHSPSP	2.12	2.22	1.62	3.00	0.87	2.93	2.06	2.22	1.80	2.79	1.52	1.99								
TOTALNH	18,424	3,827	3,830	3,555	3,665	3,547	18,443	3,832	3,816	3,552	3,693	3,550								
TOTALNHSP	97.88	97.78	98.38	97.00	99.13	97.67	97.94	97.78	98.20	97.21	98.48	98.01								
WHITENH	12,841	3,378	1,628	2,860	2,293	2,682	12,827	3,401	1,610	2,850	2,207	2,699								
WHITENHSP	68.22	86.31	41.82	78.04	62.02	73.40	68.12	86.78	41.43	78.00	60.45	74.52								
BLACKNH	5,389	400	2,139	666	1,349	835	5,420	388	2,152	670	1,380	824								
BLACKNHSP	28.63	10.22	54.94	18.17	36.49	22.85	28.78	9.90	55.38	18.50	36.80	22.75								
AIANNH	103	32	36	19	11	15	112	36	27	16	26	17								
AIANNHSP	0.55	0.82	0.67	0.52	0.30	0.41	0.59	0.66	0.69	0.44	0.69	0.47								
ASIANNH	47	14	16	6	7	4	51	11	18	5	15	2								
ASIANNHSP	0.25	0.36	0.41	0.16	0.19	0.11	0.27	0.28	0.46	0.14	0.40	0.06								
HPNH	3	2	0	0	0	1	0	0	0	0	0	0								
HPNHSP	0.02	0.05	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00								
OTHERNH	9	1	5	1	1	1	18	3	5	2	3	5								
OTHERNHSP	0.05	0.03	0.13	0.03	0.03	0.03	0.10	0.08	0.13	0.05	0.08	0.14								
MLTMNH	32	0	16	3	4	9	15	3	4	3	2	3								
MLTMNHSP	0.17	0.00	0.41	0.08	0.11	0.25	0.08	0.08	0.10	0.08	0.05	0.08								
HSP18	215	47	34	63	16	55	226	53	46	62	29	36								
HSP18P	1.55	1.69	1.20	3.25	0.58	2.01	1.62	1.90	1.62	3.22	1.05	1.32								
NONHSP18	13,678	2,733	2,792	2,736	2,739	2,678	13,683	2,735	2,787	2,734	2,744	2,683								
NONHSP18P	98.45	98.31	98.80	97.75	99.42	97.99	98.38	98.10	98.38	97.78	98.95	98.68								
WHITENH18	9,747	2,428	1,278	2,219	1,755	2,057	9,739	2,456	1,265	2,207	1,734	2,076								
WHITENH18P	70.16	87.70	45.22	79.28	63.70	75.27	70.01	88.00	44.65	78.93	62.53	76.15								
BLACKNH18	3,790	261	1,471	498	965	595	3,800	248	1,485	504	977	586								
BLACKNH18P	27.28	9.39	52.05	17.79	35.03	21.77	27.32	8.90	52.42	18.03	35.23	21.55								
AIANNH18	79	23	21	13	9	13	82	22	18	13	16	13								
AIANNH18P	0.57	0.83	0.74	0.46	0.33	0.48	0.59	0.79	0.64	0.46	0.58	0.48								
ASIANNH18	35	8	13	4	6	4	36	4	12	5	13	2								
ASIANNH18P	0.25	0.29	0.46	0.14	0.22	0.15	0.26	0.14	0.42	0.18	0.47	0.07								
HPNH18	3	2	0	0	0	1	0	0	0	0	0	0								
HPNH18P	0.02	0.07	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00								
OTHERNH18	4	1	1	1	0	1	14	3	3	2	3	3								
OTHERNH18P	0.03	0.04	0.04	0.04	0.00	0.04	0.10	0.11	0.11	0.07	0.11	0.11								
MLTMNH18	29	0	8	1	4	7	13	2	4	3	1	3								
MLTMNH18P	0.14	0.00	0.28	0.04	0.15	0.26	0.09	0.07	0.14	0.11	0.04	0.11								

$$2010 \text{ Census IDEAL POPULATION} = \frac{1,609}{4} = 402.25$$

$$TDA \text{ IDEAL POPULATION} = \frac{1,617}{4} = 404.25$$

Demographics	2010 Census, SFI (PL 94-171)									
	Comms & Percentages POST-2010 Plan				Comms & Percentages Run A of the TDA					
	Tylertown	01	02	03	04	Tylertown	01	02	03	04
DIST-ID										
TOTAL	1,609	405	399	391	414	1,617	388	411	401	407
DEV		2.8	-3.2	-11.2	11.8		-6.2	6.8	-3.2	2.8
DEVP		0.68	-0.81	-2.88	2.84		-1.57	1.64	-0.81	0.68
TOTAL18	1,333	327	320	313	273	1,344	323	335	312	274
TOTALHSP	42	12	7	9	14	45	12	11	18	4
TOTALHSPSP	2.61	2.96	1.75	2.30	3.38	2.78	3.02	2.68	4.49	0.98
TOTALNH	1,567	393	392	382	400	1,572	386	400	383	403
TOTALNHP	97.39	97.04	98.25	97.70	96.62	97.22	96.98	97.32	95.51	99.02
WHITEH	860	371	215	246	28	850	368	207	244	31
WHITEHNP	53.45	91.60	53.88	62.92	6.76	56.57	92.46	50.36	60.85	7.62
BLACKNH	679	17	174	119	369	676	14	171	122	369
BLACKHNP	42.30	4.20	43.61	30.43	89.13	41.81	3.52	41.61	30.42	90.66
AIANNH	14	5	3	3	3	19	0	12	5	2
AIANNHP	0.87	1.23	0.75	0.77	0.72	1.18	0.00	2.92	1.25	0.49
ASIANH	12	0	0	12	0	14	2	6	6	0
ASIANHNP	0.75	0.00	0.00	3.07	0.00	0.87	0.50	1.46	1.50	0.00
HPINH	0	0	0	0	0	0	0	0	0	0
HPINHNP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OTHERNH	0	0	0	0	0	5	2	1	1	1
OTHERHNP	0.00	0.00	0.00	0.00	0.00	0.31	0.50	0.24	0.25	0.25
MLTMNH	2	0	0	2	0	8	0	3	5	0
MLTMHNP	0.12	0.00	0.00	0.51	0.00	0.49	0.00	0.73	1.25	0.00
HSP18	27	7	4	8	8	26	9	5	8	4
HSP18P	2.19	2.14	1.25	2.56	2.93	2.09	2.79	1.49	2.56	1.46
NONHSP18	1,206	320	316	305	265	1,218	314	330	304	270
NONHSP18P	97.81	97.86	98.75	97.44	97.07	97.91	97.21	98.51	97.44	98.54
WHITEH18	723	302	188	210	23	717	301	183	208	25
WHITEH18P	58.64	92.35	58.75	67.09	8.42	57.64	93.19	54.63	66.67	9.12
BLACKNH18	462	14	127	81	240	464	9	132	81	242
BLACKNH18P	37.47	4.28	39.69	35.88	87.91	37.30	2.79	39.40	35.96	88.32
AIANNH18	10	4	1	3	2	11	0	6	3	2
AIANNH18P	0.81	1.22	0.31	0.96	0.73	0.88	0.00	1.79	0.96	0.73
ASIANH18	10	0	0	10	0	14	2	6	6	0
ASIANH18P	0.81	0.00	0.00	3.19	0.00	1.13	0.62	1.79	1.92	0.00
HPINH18	0	0	0	0	0	0	0	0	0	0
HPINH18P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OTHERNH18	0	0	0	0	0	5	2	1	1	1
OTHERNH18P	0.00	0.00	0.00	0.00	0.00	0.40	0.62	0.30	0.32	0.36
MLTMNH18	1	0	1	0	0	7	0	2	5	0
MLTMNH18P	0.08	0.00	0.00	0.32	0.00	0.56	0.00	0.60	1.00	0.00

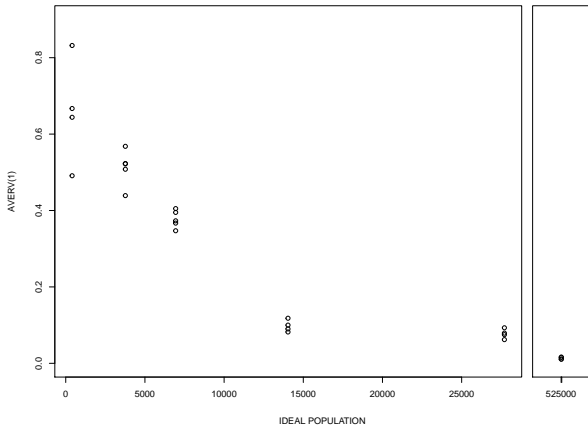
Run $i$	$C_{Ti}(g)$	$(C_{Ti}(g) - C_T(g))^2$	$(C_{Ti}(g) - C_S(g))^2$
1.	17,628	$(17,628 - 17,684.80)^2 = 3,226.24$	$(17,628 - 17,705)^2 = 5,929$
2.	17,685	$(17,685 - 17,684.80)^2 = 0.04$	$(17,685 - 17,705)^2 = 400$
3.	17,671	$(17,671 - 17,684.80)^2 = 190.44$	$(17,671 - 17,705)^2 = 1,156$
4.	17,669	$(17,669 - 17,684.80)^2 = 249.64$	$(17,669 - 17,705)^2 = 1,296$
5.	17,713	$(17,713 - 17,684.80)^2 = 795.24$	$(17,713 - 17,705)^2 = 64$
6.	17,692	$(17,692 - 17,684.80)^2 = 51.84$	$(17,692 - 17,705)^2 = 169$
7.	17,692	$(17,692 - 17,684.80)^2 = 51.84$	$(17,692 - 17,705)^2 = 169$
8.	17,640	$(17,640 - 17,684.80)^2 = 2,007.04$	$(17,640 - 17,705)^2 = 4,225$
9.	17,715	$(17,715 - 17,684.80)^2 = 912.04$	$(17,715 - 17,705)^2 = 100$
10.	17,625	$(17,625 - 17,684.80)^2 = 3,576.04$	$(17,625 - 17,705)^2 = 6,400$
11.	17,718	$(17,718 - 17,684.80)^2 = 1,102.24$	$(17,718 - 17,705)^2 = 169$
12.	17,707	$(17,707 - 17,684.80)^2 = 492.84$	$(17,707 - 17,705)^2 = 4$
13.	17,703	$(17,703 - 17,684.80)^2 = 331.24$	$(17,703 - 17,705)^2 = 4$
14.	17,649	$(17,649 - 17,684.80)^2 = 1,281.64$	$(17,649 - 17,705)^2 = 3,136$
15.	17,692	$(17,692 - 17,684.80)^2 = 51.84$	$(17,692 - 17,705)^2 = 169$
16.	17,736	$(17,736 - 17,684.80)^2 = 2,621.44$	$(17,736 - 17,705)^2 = 961$
17.	17,654	$(17,654 - 17,684.80)^2 = 948.64$	$(17,654 - 17,705)^2 = 2,601$
18.	17,684	$(17,684 - 17,684.80)^2 = 0.64$	$(17,684 - 17,705)^2 = 441$
19.	17,750	$(17,750 - 17,684.80)^2 = 4,251.04$	$(17,750 - 17,705)^2 = 2,025$
20.	17,678	$(17,678 - 17,684.80)^2 = 46.24$	$(17,678 - 17,705)^2 = 729$
21.	17,633	$(17,633 - 17,684.80)^2 = 2,683.24$	$(17,633 - 17,705)^2 = 5,184$
22.	17,720	$(17,720 - 17,684.80)^2 = 1,239.04$	$(17,720 - 17,705)^2 = 225$
23.	17,669	$(17,669 - 17,684.80)^2 = 249.64$	$(17,669 - 17,705)^2 = 1,296$
24.	17,723	$(17,723 - 17,684.80)^2 = 1,459.24$	$(17,723 - 17,705)^2 = 324$
25.	17,674	$(17,674 - 17,684.80)^2 = 116.64$	$(17,674 - 17,705)^2 = 961$
<b>Totals</b>	<b>442,120</b>	<b>27,936.00</b>	<b>38,137.00</b>

$C_T(g) = \frac{442,120}{25} = 17,684.80 \approx \mathbf{17,685}$	$C_S(g) = \mathbf{17,705}$
$\sqrt{V(1)_g} = \sqrt{\frac{27,936}{25}} = 33.43 \approx \mathbf{33}$	$\sqrt{V(2)_g} = \sqrt{\frac{38,137}{25}} = 39.06 \approx \mathbf{39}$
$RV(1)_g = \frac{\sqrt{V(1)_g}}{C_T(g)} = 0.00189 \approx \mathbf{0.002}$	$RV(2)_g = \frac{\sqrt{V(2)_g}}{C_S(g)} = 0.00221 \approx \mathbf{0.002}$

Figure 1

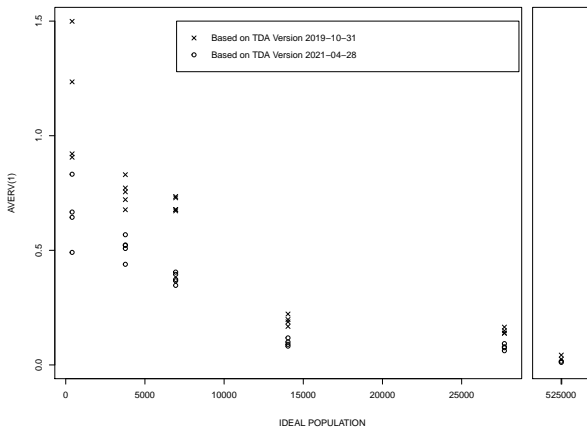
Jurisdiction	District	IDEAL POPULATION	A <sub>ERV</sub> (1)
Rhode Island	CD-01	526,283.50	0.011
Rhode Island	CD-02	526,283.50	0.016
Rhode Island	SLDU-01	27,699.10	0.062
Rhode Island	SLDU-02	27,699.10	0.093
Rhode Island	SLDU-03	27,699.10	0.079
Rhode Island	SLDU-04	27,699.10	0.075
Rhode Island	SLDL-01	14,034.2	0.118
Rhode Island	SLDL-02	14,034.20	0.082
Rhode Island	SLDL-03	14,034.20	0.090
Rhode Island	SLDL-04	14,034.20	0.100
Panola County, MS	D-01	6,941.40	0.373
Panola County, MS	D-02	6,941.40	0.405
Panola County, MS	D-03	6,941.40	0.347
Panola County, MS	D-04	6,941.40	0.395
Panola County, MS	D-05	6,941.40	0.367
Tate County Schools, MS	D-01	3,764.60	0.439
Tate County Schools, MS	D-02	3,764.60	0.508
Tate County Schools, MS	D-03	3,764.60	0.522
Tate County Schools, MS	D-04	3,764.60	0.523
Tate County Schools, MS	D-05	3,764.60	0.568
Tylertown, MS	D-01	402.25	0.667
Tylertown, MS	D-02	402.25	0.644
Tylertown, MS	D-03	402.25	0.491
Tylertown, MS	D-04	402.25	0.832

Plot of AVERV(1) for IDEAL POPULATION Values Noted Above



## II.8. CONCLUDING REMARKS FOR PART II

Figure 2



THANK YOU!

*SlidesblockgroupsDOJ.tex*

**IRC\_01321**

36/37



# Determining the Privacy-loss Budget

## Research into Alternatives to Differential Privacy

**Michael Hawes and Rolando Rodríguez**  
U.S. Census Bureau

June 4, 2021

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
**2020**

**IRC\_01322**

# Acknowledgements

**This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, from the following Census Bureau divisions and outside organizations:** ADCOM, ADDC, ADRM, CED, CEDDA, CEDSCI, CES, CSRM, DCMD, DITD, ESMD, GEO, POP, TAB, CDF, Econometrica Inc., Galois, Knexus Research Corp, MITRE, NLT, TI, and Tumult Labs.

**We also acknowledge and greatly appreciate the ongoing feedback we have received from external stakeholder groups that has contributed to the design and improvement of the Disclosure Avoidance System.**

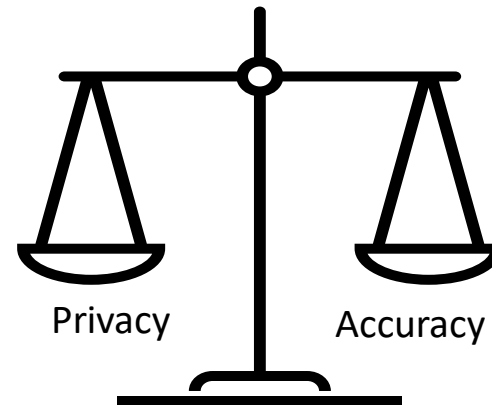
**For more information and technical details relating to the issues discussed in these slides, please contact the author at [michael.b.hawes@census.gov](mailto:michael.b.hawes@census.gov).**

**Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.**

Shape  
your future  
START HERE >

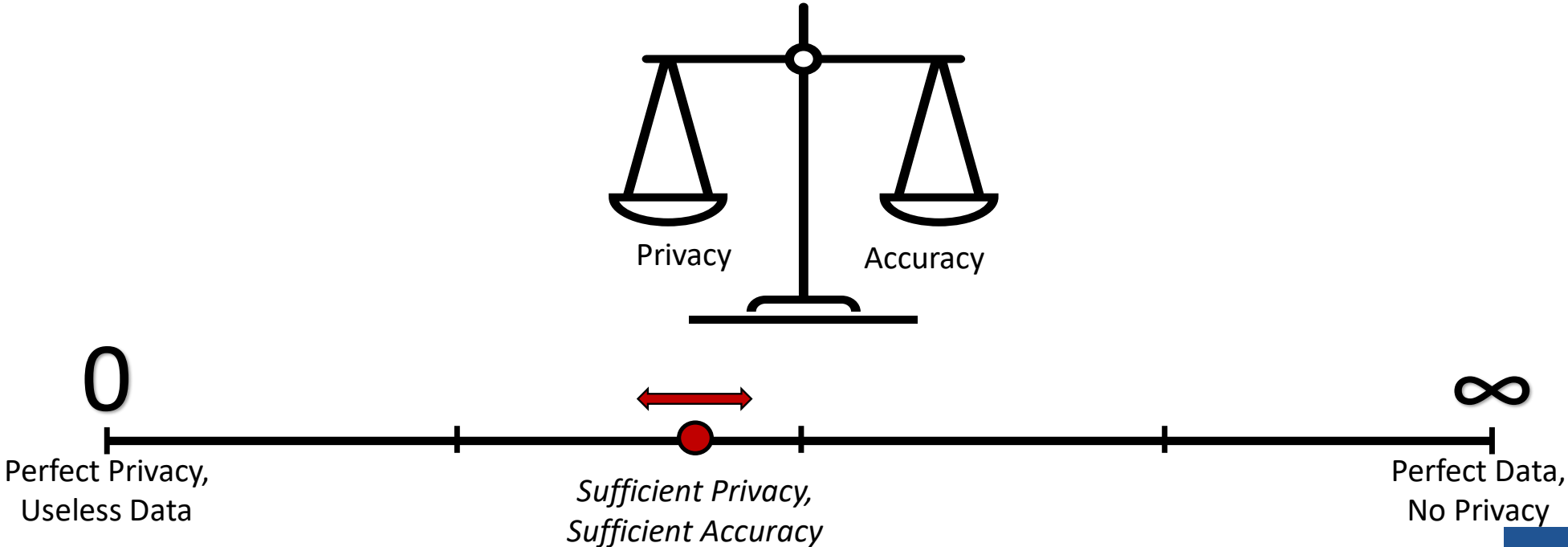
United States<sup>®</sup>  
**Census**  
**2020**

# What is a Privacy-loss Budget?



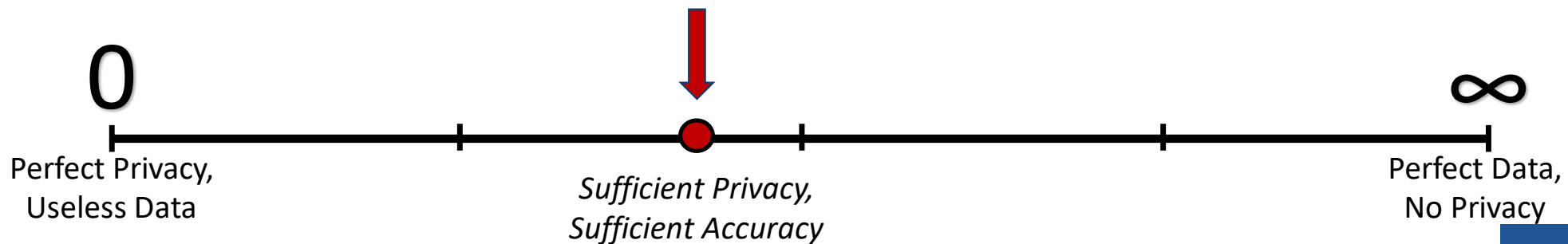
Any disclosure avoidance mechanism imposes a fundamental tradeoff between data protection (privacy/confidentiality) and data accuracy/fitness-for-use.

# What is a Privacy-loss Budget?



# What is a Privacy-loss Budget?

## Privacy-loss Budget (PLB, " $\epsilon$ ", " $\rho$ ")



2020CENSUS.GOV

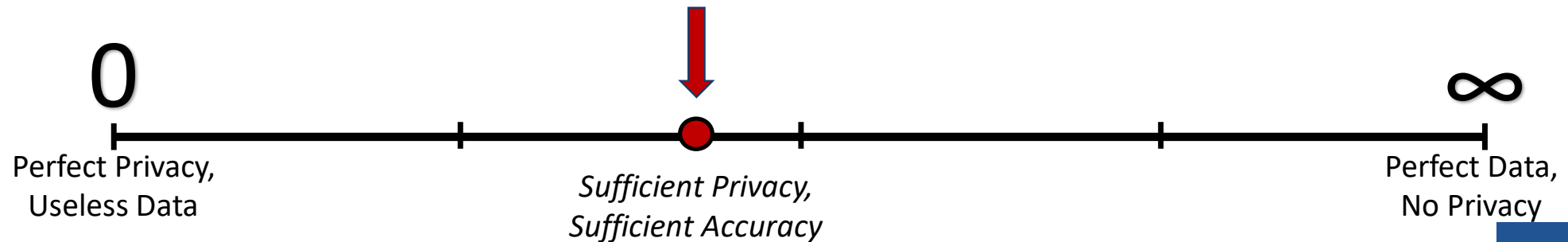
Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

IRC\_01326

# What is a Privacy-loss Budget?

Determining the optimal PLB is a (difficult) policy decision



2020CENSUS.GOV

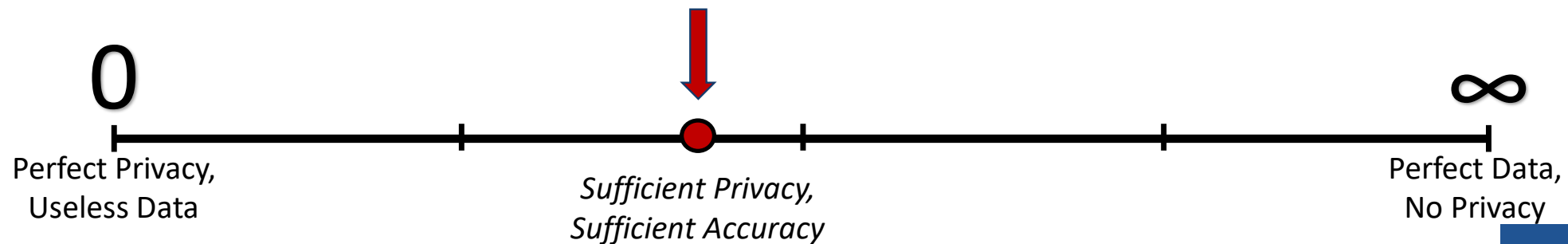
Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

IRC\_01327

# What is a Privacy-loss Budget?

Comparisons to alternative methodologies can help put these trade-offs into perspective



2020CENSUS.GOV

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Background

**DAS Reconstruction Team efforts since February 2020**

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
**2020**



# Formation and goals of DAS Reconstruction group

- The DAS Science and DevOps team continue to finalize implementation of the TopDown Algorithm for 2020 Census production
- In February 2020, a group in CED-DA began assessing the potential impacts of swapping, using an algorithm based upon the one used for the 2010 Census
- This team has become the DAS Reconstruction team, and has since performed these swapping experiments and generated preliminary assessment of the impact of suppression

# Suppression

Experiments based upon 1980 Census suppression rules and OMB race categories

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Suppression Primer

- Suppression involves removing information from published tables to protect privacy
- The 1980 Census used two types of suppression: table suppression and cell suppression
- Table suppression involves deleting tables that fail specified thresholds
- Cell suppression involves deleting individual table cells that fail specific thresholds
- Cell suppression is typically harder to implement due to the need for complimentary suppression

# Suppression Primer: Complementary Cell Suppression

Variable A	Category 1	Category 2	
Variable B			
Category 1	20	17	37
Category 2	2	15	17
	22	32	54

Cell value is too small

Variable A	Category 1	Category 2	
Variable B			
Category 1	20	17	37
Category 2	S	15	17
	22	32	54

Suppress the value

# Suppression Primer: Complementary Cell Suppression

Variable A	Category 1	Category 2	
Variable B			
Category 1	20	17	37
Category 2	S	15	17
	22	32	54

Other cells and table margins allow recovery of suppressed value

Variable A	Category 1	Category 2	
Variable B			
Category 1	S	S	37
Category 2	S	S	17
	22	32	54

Complementary suppression prevents this from happening

# Suppression from the 1980 Census

- The DAS Reconstruction team assessed the impact of applying 1980 Census-based suppression rules to the P.L. 94-171 (redistricting data) and Summary File 1 products (the “Demographic and Housing Characteristics” (DHC) file in 2020) based on the 2010 Census Edited File (CEF)
- The team used race and ethnicity categories specified by the Office of Management and Budget in Statistical Policy Directive 15 (1997) and implemented by the Department of Justice Voting Section
  - White alone
  - Black alone or in combination with white
  - Asian alone or in combination with white
  - Native Hawaiian or other Pacific Islander alone or in combination with white
  - American Indian or Alaska Native alone or in combination with white
  - Some other race alone or in combination with white
  - Two or more races, except as explicitly noted in the categories above
  - Hispanic/Not-Hispanic

# Suppression from the 1980 Census

## P.L. 94-171 Redistricting Data

- Table Suppression: Whole tables were suppressed (not published) for geographies with between 1 and 14 persons in any of the race/ethnicity groups
  - Applied to two tables:
    - (P3) Race for the Population 18 Years and Over, and
    - (P4) Hispanic or Latino, and not Hispanic or Latino, by Race for the Population 18 Years and Over
- Cell Suppression: Cell counts of 1 or 2 were replaced by 0
  - Applied to two tables:
    - (P1) Race
    - (P2) Hispanic or Latino, and not Hispanic or Latino by Race

## Additional Summary File (SF1) Data

- Table Suppression: Whole tables that are not dedicated solely to race and ethnicity are suppressed if their geographies have between 1 and 14 persons.
- For all person-level tables

# Impact of Suppression Rules on Privacy Risk

- Suppression, if done correctly, removes information from the tables that are released
- This means that enough suppression done on a set of tables can prevent re-identification attacks based on reconstruction of microdata from those tables
- While this would eliminate the risk of a specific attack on a specific set of tables, it is not equivalent to the broad privacy protection associated with formal privacy definitions



# Suppression Results: P.L. 94-171

- Under the 1980 suppression rules, tables P1 and P2 would have cell suppression applied only
- Cells with counts of 1 or 2 would be reported as 0
- The population total margin of P1 and P2 is never suppressed
- *These results include only primary cell suppressions*
- *Complementary suppressions would be necessary to prevent recovering cell values from margins*

2020CENSUS.GOV

## P1: Race

Geography	Total Cells	Cells Changed to Zero	% Cells Changed
Nation	7	0	0
State	357	0	0
County	22,001	530	2.4
Tract	507,717	28,024	5.5
Block Group	1,518,048	153,914	10.1
Block	43,449,189	3,538,888	8.1

DRB clearance number CBDRB-FY21-213

## P2: Hispanic or Latino, and Not Hispanic or Latino by Race

Geography	Total Cells	Cells Changed to Zero	% Cells Changed
Nation	14	0	0
State	714	0	0
County	44,002	2,987	6.8
Tract	1,015,434	110,081	10.8
Block Group	3,036,096	440,539	14.5
Block	86,898,378	5,071,570	5.8

DRB clearance number CBDRB-FY21-213

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Suppression Results: P.L. 94-171

- Results of the experiment show that table suppression for P.L. 94-171 tables P3 and P4 would exceed 84% and 87% (respectively) for on-spine geographies below the county level (tract, block group, block)

2020CENSUS.GOV

## P3: Race For The Population 18 Years and Over

Geography	Total Tables	Suppressed Tables	% Tables Suppressed
Nation	1	0	0
State	51	0	0
County	3,143	1,610	51.2
Tract	72,531	61,177	84.3
Block Group	216,864	207,643	95.7
Block	6,206,505	5,204,047	83.8

DRB clearance number CBDRB-FY21-213

## P4: Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over

Geography	Total Tables	Suppressed Tables	% Tables Suppressed
Nation	1	0	0
State	51	0	0
County	3,143	2,645	84.2
Tract	72,531	72,346	99.7
Block Group	216,864	216,759	100.0
Block	6,206,505	5,445,153	87.7

DRB clearance number CBDRB-FY21-213

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Suppression Results: P.L. 94-171

- The team also assessed the potential impact of cell suppression on tables P3 and P4
- This would imply adding voting age as part of the cell suppression criteria
- *These results include only primary cell suppressions*
- *Complementary suppressions would also be necessary to prevent recovering cell values from margins*

2020CENSUS.GOV

## P3: Race For The Population 18 Years and Over

Geography	Total Cells	Cells Changed to Zero	% Cells Changed
Nation	7	0	0
State	357	0	0
County	22,001	822	3.7
Tract	507,717	38,439	7.6
Block Group	1,518,048	204,853	13.5
Block	43,449,189	4,200,018	9.7

DRB clearance number CBDRB-FY21-213

## P4: Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over

Geography	Total Cells	Cells Changed to Zero	% Cells Changed
Nation	14	0	0
State	714	0	0
County	44,002	4,078	9.3
Tract	1,015,434	146,400	14.4
Block Group	3,036,096	533,314	17.6
Block	86,898,378	5,822,712	6.7

DRB clearance number CBDRB-FY21-213

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Suppression Results: SF1

- The team assessed the impact of table suppression on additional 2010 SF1 tables by counting how many geographies meet broad restrictions on the total population and housing units
- This assessment showed that suppression of SF1 at the block level would exceed 38% for person-level tables and 32% for housing unit tables
- Additional SF1 table suppressions would be necessary at the block group and tract levels as well

2020CENSUS.GOV

## SF1: Geographies meeting criteria for person table suppression

Geography	Total populated	Population meets criteria	% Meets Criteria
Nation	1	0	0
State	51	0	0
County	3,143	0	0
Tract	72,531	131	0.2
Block Group	216,864	204	0.1
Block	6,207,027	2,401,802	38.7

DRB clearance number CBDRB-FY21-213

## SF1: Geographies meeting criteria for housing table suppression

Geography	Total occupied	Housing unit count meets criteria	% Meets Criteria
Nation	1	0	0
State	51	0	0
County	3,143	0	0
Tract	72,425	182	0.3
Block Group	216,598	307	0.1
Block	6,188,078	2,027,988	32.8

DRB clearance number CBDRB-FY21-213

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Swapping

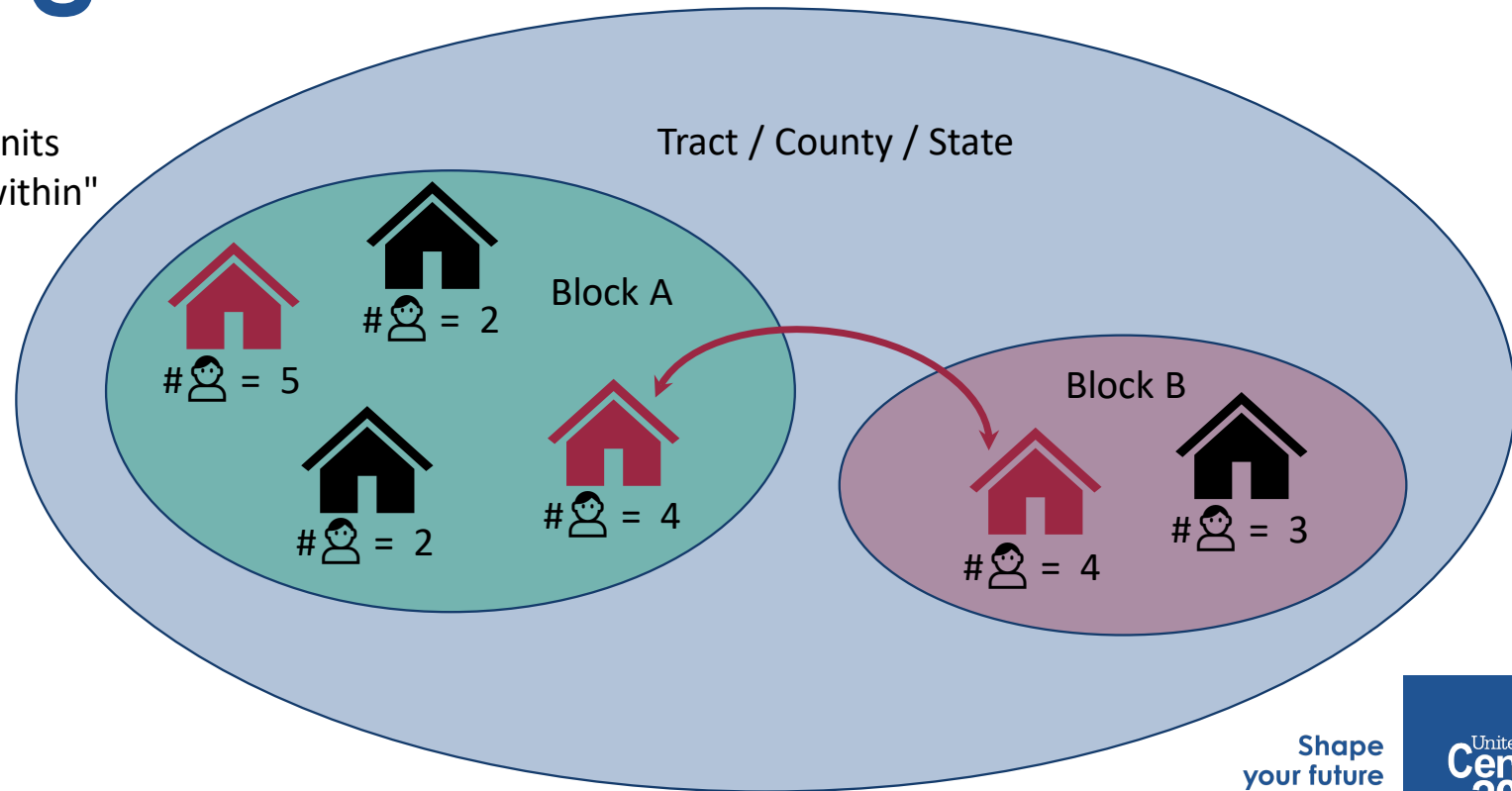
Relaxations and extensions of the 2010 Census swapping algorithm

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Swapping Primer

1. Determine key to match units
2. Choose "between" and "within" geographies
3. Determine units to swap
4. Select swap rate
5. Find swap pairs



2020CENSUS.GOV

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

IRC\_01343

# Adapting the 2010 Swapping Algorithm for higher rates

- Initial efforts of the DAS Reconstruction team focused on adapting the 2010 Census swapping to support higher swap rates, up to 100% if necessary
- This algorithm now has the following parameters and adjustments:
  - The desired swap rate
  - The list of invariants (the swap “key”)
  - Mechanisms for relaxing invariants and extending swapping beyond tracts

# Swapping Experiments

- The DAS Reconstruction team has prepared swapped files for numerous iterations of the parameters
  - Swap rates ranging from 5% to 50% of housing units
  - Pre-swap perturbation of household size by  $\pm 1$  for up to 80% of housing units
  - Pre-swap perturbation of tract within county or within state for up to 70% of housing units
- At the beginning of CY2021, the team began to assess the impact of these parameters on the outcomes of the reconstruction-abetted re-identification attack on the 2010 Census



# Swapping Results

- The key swapping outcomes of those experiments have been:
  - Low swap rates have essentially no impact on re-identification outcomes; they are essentially the same as for the 2010 SF1
  - High swap rates have only a minimal impact on re-identification outcomes, with accuracy metrics inferior to the 4/28/2021 Disclosure Avoidance System (DAS) Privacy-Protected Microdata File (PPMF)
- These imply that middling swap rates, as implemented, may match the TopDown Algorithm in terms of accuracy but will have a low impact on reducing re-identification

Swap Parameters				Reidentification		
Experiment	Swap %	%HH Size Perturbed	%Tract perturbed	Putative % of Population	Confirmed % of Population	Precision (Confirmed/Putative)
2010 HDF	-	0	-	44.60	16.85	37.79
SwapLow	5	0	0	44.38	16.52	37.23
SwapHigh	50	50	70	42.69	12.96	30.37

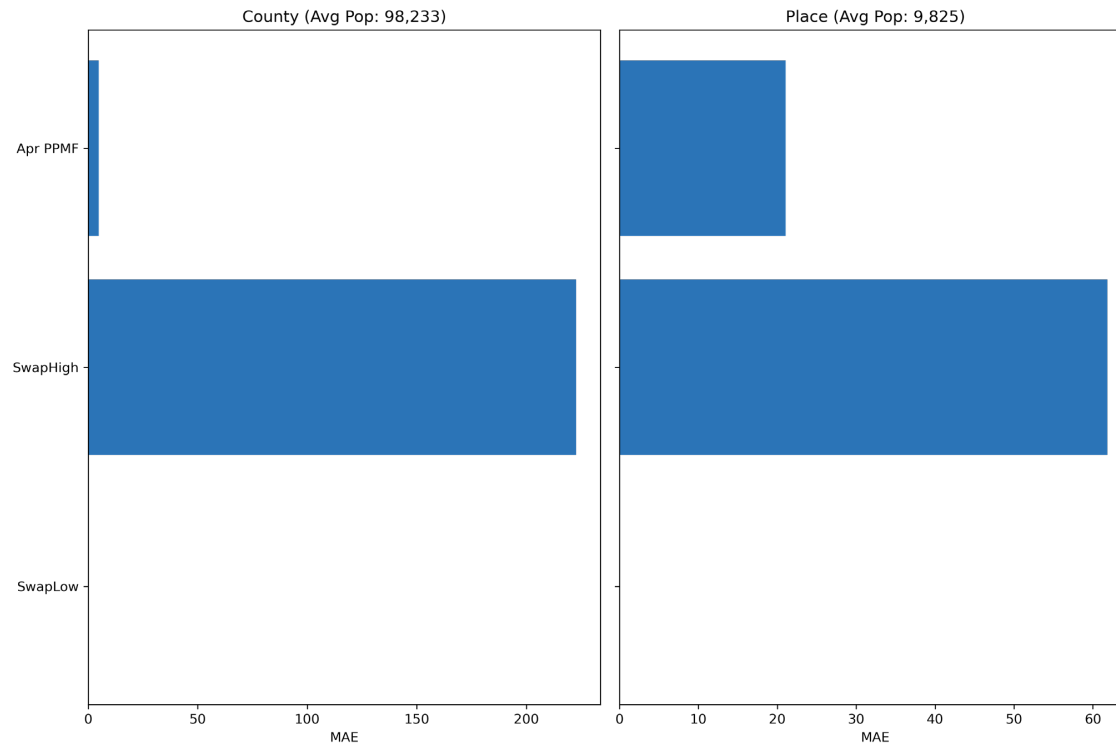
2020CENSUS.GOV  
DRB clearance number CBDRB-FY21-213

[START HERE >](#)

United States<sup>®</sup>  
**Census**  
**2020**

# Swapping Results

Comparison of mean absolute error (MAE) for total population for county and incorporated place size categories



DRB clearance number CBDRB-FY21-213

2020CENSUS.GOV

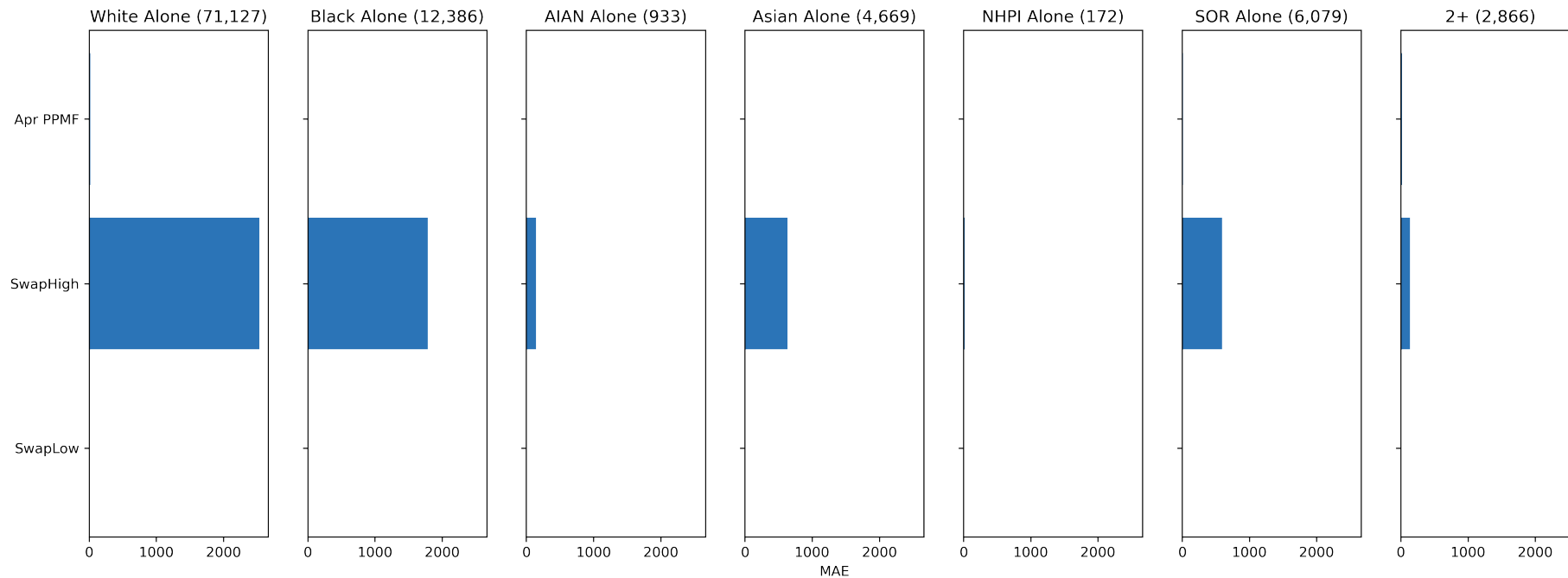
Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

IRC\_01347

# Swapping Results

Comparison of mean absolute error (MAE) for race alone for counties



DRB clearance number CBDRB-FY21-213

2020CENSUS.GOV

Shape  
your future  
START HERE >

United States  
Census  
2020

# Final Considerations

- None of the algorithms described herein adheres to a formal definition or semantic for privacy loss, and they are only being assessed against one attack strategy (the 2010 Census reconstruction-abetted re-identification attack)
- Implementation of the 1980 Census suppression rules would lead to extreme amounts of table suppression for sub-state on-spine (county, tract, block group, block) geographies
- Implementation of relaxations and extensions of the 2010 Census swapping algorithm would yield little improvement in re-identification outcomes even at high swap rates
- Production implementation of either suppression or swapping is expected to take at least an additional 6 months after a decision to implement them

Stay Informed:  
Subscribe to the 2020 Census Data  
Products Newsletters

\*Search “Disclosure Avoidance” at [www.census.gov](http://www.census.gov)



## 2020 Census Data Products Newsletters

Sign up for news and information about 2020 Census Data Products and the implementation of the new Disclosure Avoidance System.

[SIGN-UP FOR NEWSLETTERS](#)

### Past Issues:

May 04, 2021

**Webinar Today (5/4): Differential Privacy 101**

April 30, 2021

**Save the Dates for Additional Webinars Throughout May**

April 28, 2021

**New DAS Update Meets or Exceeds Redistricting Accuracy Targets**

April 19, 2021

**New Demonstration Data Will Feature Higher Privacy-loss Budget**

April 07, 2021

**Meeting Redistricting Data Requirements: Accuracy Targets**

February 23, 2021

**The Road Ahead: Upcoming Disclosure Avoidance System Milestones**

[START HERE >](#)

Stay Informed:  
Visit Our Website

\*Search “Disclosure Avoidance” at [www.census.gov](http://www.census.gov)

**“Disclosure Avoidance Webinar Series:  
view archived presentations”**



## 2020 Census Data Products: Disclosure Avoidance Modernization

Modern computers and today's data-rich world have rendered the Census Bureau's traditional confidentiality protection methods obsolete. Those legacy methods are no match for hackers aiming to piece together the identities of the people and businesses behind published data.

A powerful new disclosure avoidance system (DAS) designed to withstand modern re-identification threats will protect 2020 Census data products (other than the apportionment data; those state-level totals remain unaltered by statistical noise).

Inspired by cryptographic principles, the 2020 DAS is the only solution that can respond to this threat while maximizing the availability and utility of published census data.

### Learn More:

- \*\* Disclosure Avoidance Webinar Series: Join live or view archived presentations \*\*
- Census Bureau Declarations for Alabama v. Commerce II Litigation [4.2 MB]
- Video Presentation: Differential Privacy and the 2020 Census [242 MB]
- Animation: Protecting Privacy with Math, a collaboration with MinutePhysics
- Infographic: A History of Census Privacy Protections
- JASON report on Privacy Methods for the 2020 Census
- All Disclosure Avoidance Working Papers

### Latest Updates

- Disclosure Avoidance System Development

### Data Products Newsletter

April 30, 2021  
Save the Dates for Additional Webinars Throughout May

[SIGN-UP FOR NEWSLETTERS](#) [VIEW ALL NEWSLETTERS](#)

# Questions?



We only use cookies that are necessary for this site to function, and to provide you with the best experience. Learn more in our [Cookie Statement](#). By continuing to use this site, you consent to the use of cookies.



## Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results



**JUNE 9, 2021** — The U.S. Census Bureau’s [Data Stewardship Executive Policy Committee](#) (DSEP) announced it has selected the settings and parameters for the Disclosure Avoidance System (DAS) for the 2020 Census redistricting data (PL-94-171). The DAS uses a mathematical algorithm to ensure that the privacy of individuals is sufficiently protected while maintaining high levels of accuracy in the statistics we produce.

The Census Bureau released the [first “beta” version](#) of the DAS in October 2019, and released further demonstration data products in May, September, and November 2020, and in April 2021. During this process, independent experts and stakeholders, along with data users, have provided extensive feedback to help shape each subsequent test product and to inform the decisions.

After reviewing feedback from the data user community regarding the [April 2021 demonstration data](#), the committee approved a revised algorithm that makes notable improvements in the accuracy of the population counts for places, Minor Civil Divisions, American Indian and Alaska Native tribal areas, and for race and ethnicity statistics, and ensures the accuracy of data necessary for redistricting and Voting Rights Act enforcement.



The approved DAS production settings reflect a total privacy-loss budget for the redistricting data product (represented by “ $\epsilon$ ,” the Greek letter “epsilon”) of  $\epsilon=19.61$ , which includes  $\epsilon=17.14$  for the persons file and  $\epsilon=2.47$  for the housing unit data. The increased privacy-loss budget over the levels reflected in the April 2021 demonstration data—which will lead to lower noise infusion than that in the April 2021 demonstration data—was primarily allocated to the total population and race by ethnicity queries at the block group level and above.

Our Disclosure Avoidance team will use these parameters to prepare the TopDown Algorithm for final system integration testing in anticipation of the DAS application phase of our data processing and related quality assurance checks that will begin later this month. The data will be run and quality checked multiple times prior to release, which are yet further steps in the process that will culminate in the states receiving the final redistricting numbers by August 16.

“The decisions strike the best balance between the need to release detailed, usable statistics from the 2020 Census with our statutory responsibility to protect the privacy of individuals’ data,” said Ron Jarmin, acting director of the U.S. Census Bureau. “They were made after many years of research and candid feedback from data users and outside experts – whom we thank for their invaluable input.”

The 2020 DAS algorithm injects carefully calibrated statistical “noise” to obscure individual data responses. The 2010 and other recent censuses also injected statistical noise into the data, but in a less precise and more ad hoc manner, primarily using a data-swapping methodology. Recent research has confirmed that today’s superior computational technologies have rendered the methods used in 2010 and earlier censuses ineffective against reidentification attacks. The Census Bureau’s recent blog, [Modernizing Privacy Protections for the 2020 Census: Next Steps](#), discusses the privacy challenges that led to the change.

The chosen global privacy-loss budget of  $\epsilon=19.61$  is exponentially higher than the  $\epsilon=12.2$  budget used in the April 2021 demonstration data. In making its decisions, DSEP gave significant consideration to the feedback we received from our data users who analyzed the April 2021 demonstration data. That feedback, and steps taken to address those comments, include the following:

- Stakeholders identified a regression in the accuracy of data for tribal geographies and other off-spine geographies. The DAS team made changes to the ‘optimized spine’ to address these concerns; those changes were integrated into the spine that was approved by DSEP.
- Stakeholders identified several measures of bias in the summary metrics that they indicated were areas of concern. In particular, stakeholders addressed concerns about both geographic bias (i.e., the accuracy of population counts being different at larger and smaller geographies) and characteristic bias (counts of racially or ethnically diverse geographies being different than more racially or ethnically homogenous areas). The DAS team made changes to the post-processing system parameters to address these concerns; those changes were integrated into the parameters that were approved by DSEP.
- Data users identified a need for more accuracy in race and ethnicity statistics at many levels of geography. The DAS team addressed those concerns by allocating additional privacy-loss budget to the race and ethnicity queries at various levels of geography; those changes were integrated into the global privacy-loss budget and privacy-loss budget allocations that were approved by DSEP.

- Data users identified a need for more accuracy at the place, Minor Civil Division, and tract levels. The DAS team addressed these concerns both through changes to the optimized geographic spine and through allocation of privacy-loss budget; those changes were integrated into the privacy-loss budget allocations and system parameters that were approved by DSEP.
- Data users identified a need for more accurate statistics on occupancy rates at the block group and higher levels of geography. The DAS team addressed those concerns by allocating additional privacy-loss budget to the housing unit data; that change was integrated into the global privacy-loss budget and privacy-loss budget allocations that were approved by DSEP.

These improvements – as well as other adjustments to the system – were then verified against a broad suite of accuracy measures to ensure that they successfully addressed the feedback we received. We are not able to satisfy all stakeholder feedback. For example, some data users recommended nearly perfect accuracy in block-level data, which we are unable to achieve because it would undermine the ability to implement a functional disclosure avoidance system. We are both legally and ethically bound to protect the privacy of the data provided by and on behalf of our respondents.

In September, the Census Bureau anticipates releasing a final set of demonstration data that applies the privacy-loss budget and settings from today's decisions to the 2010 Census P.L. 94-171 redistricting data. Demonstration data allow data users to compare a DAS-protected version of 2010 Census results with the published 2010 Census results.

The Census Bureau will also release the DAS production code base. This is a benefit of this Census' algorithm-based system—unlike the confidential swapping methods used in previous Censuses, the 2020 DAS algorithm allows this level of transparency without risking the exposure of protected data.

Details of the settings and technical parameters for the 2020 DAS will be shared in the coming weeks. Background information is available at [census.gov](https://www.census.gov).

---

### **2021 Key Dates, Redistricting (P.L. 94-171) Data Product**

#### **June 8:**

- The Census Bureau's Data Stewardship Executive Policy (DSEP) Committee made the final determination of PLB, system parameters based on data user feedback for P.L. 94-171.

#### **Late June:**

- Final DAS production run and quality control analysis begins for P.L. 94-171 data.

#### **By August 16:**

- Release 2020 Census P.L. 94-171 data as Legacy Format Summary File\*.

#### **September:**

- Census Bureau releases PPMFs and Detailed Summary Metrics from applying the production version of the DAS to the 2010 Census data.

- Census Bureau releases production code base for P.L. 94-171 redistricting summary data file and related technical papers.

**By September 30:**

- Release 2020 Census P.L. 94-171 data\*\* and Differential Privacy Handbook.

\* Released via Census Bureau FTP site.

\*\* Released via data.census.gov.

Was this forwarded to you?

Sign up to receive your own copy!

Sign Up!

**Useful Links:**

- Research Paper: [Assessing the Reliability and Variability of the TopDown Algorithm for Redistricting Data](#)
- [Disclosure Avoidance Webinar Series](#)
- [IPUMS NHGIS Privacy-Protected Census Demonstration Data](#)
- [DAS Updates](#)
- [Progress Metrics and Data Runs](#)
- [Newsletter Archives](#)
- [All DAS FAQs](#)

Contact Us

**About Disclosure Avoidance Modernization**

The Census Bureau is protecting 2020 Census data products with a powerful new cryptography-based disclosure avoidance system known as "differential privacy." We are committed to producing 2020 Census data products that are of the same high quality you've come to expect while protecting respondent confidentiality from emerging privacy threats in today's digital world.

Share This

**Stay connected with us!**

Join the conversation on social media.



SUBSCRIBER SERVICES:

[Subscriber Settings](#) | [Remove me from All Subscriptions](#) | [Help](#)

**Get Email Updates**

Email Address  e.g. name@example.com

<https://content.govdelivery.com/accounts/USCENSUS/bulletins/2e32ea9>