# 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.

No. 3:21-CV-211-RAH-ECM-KCN

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.

# PLAINTIFFS' REPLY IN SUPPORT OF COMBINED MOTION FOR A PRELIMINARY INJUNCTION AND PETITION FOR A WRIT OF MANDAMUS

# TABLE OF CONTENTS

TABLE OF CONTENTS	i
TABLE OF EXHIBITS	ii
TABLE OF AUTHORITIES i	v
ARGUMENT	6
I. Plaintiffs Have Standing To Sue	6
A. Defendants' Application of Differential Privacy Harms Plaintiffs. This Court Can Redress These Harms	6
1. Differential Privacy Harms Plaintiffs	6
2. Defendants Are Applying Differential Privacy	6
3. The Harms Caused By Differential Privacy Are Redressable 1	8
B. Defendants' Delay in Releasing the Tabulations of Populations Harms Plaintiffs. This Court Can Redress That Harm	2
1. The Delay Harms Plaintiffs2	3
2. The Delay is Traceable to the Bureau's February 12 Decision	5
3. The Delay is Redressable	5
II. Plaintiffs Are Likely To Prevail On The Merits	7
A. Defendants' Application of Differential Privacy Is Unlawful	7
1. The Census Act Requires Accurate Tabulations of Population2	7
2. Section 209 Provides Plaintiffs With a Cause of Action	1
3. Plaintiffs Brought Their Claims Neither Too Early Nor Too Late	2
B. Defendants' Delay Releasing the Tabulations of Population Is Unlawful	8
<ol> <li>Section 209 Affords Plaintiffs a Cause of Action Because Differential Privacy Contributed to the Delay</li></ol>	8
2. Defendants' Delay Violates the Census Act	1

3. Defendants' Delay Violates the Administrative Procedure Act	. 51
III. The Equities Warrant Injunctive Relief.	. 53
IV. In The Alternative, Plaintiffs Are Entitled To A Writ Of Mandamus	. 53
CONCLUSION	. 54
CERTIFICATE OF SERVICE	. 56

# TABLE OF EXHIBITS

1.	Email from James Whitehorne to John Abowd (Sept. 30, 2020)1, 4
2.	U.S. Census Bureau, Statistical Expertise & General Research Topics (Sept. 2020), https://perma.cc/Z6JK-RLY51, 10, 11, 18, 28
3.	Simson L. Garfinkel, U.S. Census Bureau, <i>Modernizing Disclosure Avoidance:</i> <i>Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the</i> <i>2018 End-to-End Test (Continued)</i> (Sept. 15, 2017), https://perma.cc/H325-562D1, 10, 28, 29
4.	Email from John Abowd to Ron Jarmin (Jul. 7, 2020)2, 11, 30
5.	Email from John Abowd to Gary Benedetto (Jul. 28, 2020)2, 30
6.	Expert Report of Steven Ruggles passim
7.	Ron Jarmin, U.S. Census Bureau, <i>Census Bureau Adopts Cutting Edge Privacy</i> <i>Protections for 2020 Census</i> (Feb. 15, 2019), https://perma.cc/54R7-YJQS
8.	U.S. Census Bureau, <i>Invariants Set for 2020 Census Data Products</i> (Nov. 24, 2020), https://perma.cc/JET7-WDCL
9.	Supplemental Declaration of Clay S. Helms
10.	John Abowd, Tweetorial: Reconstruction-abetted re-identification attacks and other traditional vulnerabilities (Apr. 7, 2019), https://perma.cc/LUH3-QC8R
11.	Email from Benjamin Overholt to Nathaniel Cogley (Oct. 22, 2020)
12.	Defendants' Responses to Plaintiffs' First Request for Admissions
13.	U.S. Census Bureau, Census Scientific Advisory Committee Fall 2020 Meeting Recommendations (Sept. 17-18, 2020), https://perma.cc/RR3S-XEAU
14.	Email from John Abowd to Danah Boyd (Oct. 7, 2020)

# **TABLE OF AUTHORITIES**

# Cases

<ul> <li>Ala. Legis. Black Caucus v. Alabama,</li> <li>989 F. Supp. 2d 1227 (M.D. Ala. 2013), vacated and remanded on other grounds,</li> <li>575 U.S. 254 (2015)</li></ul>
Amoco Prod. Co. v. Vill. of Gambell, 480 U.S. 531 (1987)
Armstrong v. Exceptional Child Ctr., Inc., 575 U.S. 320 (2015)
<i>Baldridge v. Shapiro</i> , 455 U.S. 345 (1982)
Black Warrior Riverkeeper, Inc. v. U.S. Army Corps of Engineers, 781 F.3d 1271 (11th Cir. 2015)
Canal Auth. of Fla. v. Callaway, 489 F.2d 567 (5th Cir. 1974)
<i>City of Detroit v. Franklin</i> , 4 F.3d 1367 (6th Cir. 1993)
City of New York v. U.S. Dep't of Com., 822 F. Supp. 906 (E.D.N.Y. 1993), vacated on other grounds, 34 F.3d 1114 (2d Cir. 1994)
<i>Crawford v. Marion Cty. Election Bd.</i> , 553 U.S. 181 (2008)
<i>Darby v. Cisneros</i> , 509 U.S. 137 (1993)
Davis v. FEC, 554 U.S. 724 (2008)
<i>Dep't of Com. v. New York</i> , 139 S. Ct. 2551 (2019)
Dep't of Com. v. U.S. House of Representatives, 525 U.S. 316 (1999) passim
<i>Dep't of Homeland Sec. v. Regents of the Univ. of Cal.</i> , 140 S. Ct. 1891 (2020)

Duke Power Co. v. Envtl. Study Grp., 438 U.S. 59 (1978)	
<i>FEC v. Akins</i> , 524 U.S. 11 (1008)	7
524 0.5. 11 (1998)	
Focus on the Family v. Pinellas Suncoast Transit Auth., 344 F.3d 1263 (11th Cir. 2003)	17, 25
Franklin v. Massachusetts, 505 U.S. 788 (1992)	26, 44, 45, 51
<i>Georgia v. Evans</i> , 316 U.S. 159 (1942)	
<i>Karcher v. Daggett</i> , 462 U.S. 725 (1983)	12, 15
<i>Lujan v. Defs. of Wildlife</i> , 504 U.S. 555 (1992)	
Maine Cmty. Health Option v. United States, 140 S. Ct. 1308 (2020)	
Maryland v. King, 567 U.S. 1301 (2012)	
Muransky v. Godiva Chocolatier, Inc., 979 F.3d 917 (11th Cir. 2020)	
N.Y. Tr. Co v. Comm'r of Internal Revenue, 68 F.2d 19 (2d Cir. 1933)	
NAACP v. Bureau of the Census, 399 F. Supp. 3d 406 (D. Md.), aff'd in part, rev'd in part and remanded, 945 F.3d 183 (4th Cir. 2019)	43, 44
<i>NLRB v. Noel Canning</i> , 573 U.S. 513 (2014)	
<i>Ohio v. Raimondo</i> , No. 3:21-cv-064, 2021 WL 1118049 (S.D. Ohio Mar. 24, 2021)	
Pictet Overseas Inc. v. Helvetia Trust, 905 F.3d 1183 (11th Cir. 2018)	

Pub. Citizen v. Dep't of Just.,         491 U.S. 440 (1989)
<i>Return Mail, Inc. v. U.S. Postal Serv.</i> , 139 S.Ct. 1853 (2019)
<i>Rucho v. Common Cause</i> , 139 S. Ct. 2484 (2019)
<i>Shaw v. Reno</i> , 509 U.S. 630 (1993)
<i>Siegel v. Lepore</i> , 234 F.3d 1163 (11th Cir. 2000)
<i>Spokeo, Inc. v. Robins,</i> 136 S. Ct. 1540 (2016)
<i>Stolt-Nielsen S.A. v. AnimalFeeds Int'l Corp.</i> , 559 U.S. 662 (2010)
<i>Town of Chester v. Laroe Ests., Inc.,</i> 137 S. Ct. 1645 (2017)
<i>Trump v. New York</i> , 141 S. Ct. 530 (2020)
U.S. Army Corps of Eng'rs v. Hawkes Co., 136 S. Ct. 1807 (2016)
U.S. House of Representatives v. U.S. Dep't of Com., 11 F. Supp. 2d 76 (D.D.C. 1998)
United States v. Schmidt, 675 F.3d 1164 (8th Cir. 2012)
<i>Utah v. Evans</i> , 536 U.S. 452 (2002)
<i>Whitman v. Am. Trucking Assocs.</i> , 531 U.S. 457 (2001)
<i>Wilding v. DNC Servs. Corp.</i> , 941 F.3d 1116 (11th Cir. 2019)

Wisconsin v. City of New York, 517 U.S. 1 (1996)	
Zimmerman v. City of Austin,	
881 F.3d 378 (5th Cir. 2018)	

# Alabama Constitution

Ala. Const. art. IX, § 201	. 23
----------------------------	------

# United States Statutes and Acts of Congress

2 U.S.C. § 2a(b)	
5 U.S.C. § 551	
5 U.S.C. § 706	42, 52
7 U.S.C. § 361c(c)(2)	
7 U.S.C. § 2663(b)(4)	
13 U.S.C. § 8	passim
13 U.S.C. § 9	passim
13 U.S.C. § 141(c)	passim
28 U.S.C. § 2284	
49 U.S.C. § 5305(d)(1)(A)(i)	
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 Authority**

Andy Beveridge, Sixteen States Sue to Block Census Bureau Data Privacy Method	
(Apr. 19, 2021), https://perma.cc/WEJ9-TEXY	
Antonin Scalia & Bryan A. Garner, Reading Law:	
The Interpretation of Legal Texts (2012)	

Aref N. Dajani et al., Presentation to Census Scientific Advisory Committee, <i>The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau</i> (Sept. 2017), https://perma.cc/5GKQ-LVRR
Br. of Petrs, <i>Baldridge v. Shapiro</i> , No. 80-1436, 1981 WL 389922 (Jul. 13, 1981)
JASON, Formal Privacy Methods for the 2020 Census (Apr. 2020), https://perma.cc/G8ZM-YNN6
John M. Abowd et al., <i>The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau</i> (July 2020), https://perma.cc/634N-4XPU
<ul> <li>Margo Anderson, Baker v. Carr, the Census, and the Political and Statistical Geography of the United States: The Origin and Impact of Public Law 94-171, 62 Case W. Res. L. Rev. 1153 (2012)</li> </ul>
Reply Br. of Petrs, <i>Baldridge v. Shapiro</i> , No. 80-1436, 1981 WL 389926 (Nov. 25, 1981)
U.S. Census Bureau, Decennial Census P.L. 94-171 Redistricting Data (Mar. 15, 2021), https://perma.cc/MF5A-CMQK
U.S. Census Bureau, <i>Meeting Redistricting Data Requirements: Accuracy Targets</i> (Apr. 7, 2021), https://perma.cc/Y3LP-NVR7
U.S. Census Bureau, Summary File 1 Dataset, https://perma.cc/38JG-RAFD 19
U.S. Census Bureau, Summary File 2 Dataset, https://perma.cc/S276-KY59 19

We are not a think tank, a university, or a for profit business chasing after deep pocketed clients, we are a service to the country.... We must ensure we are providing data that can be used for the needs of all communities. I will still promote and defend what we end up producing. However, I am trying to underscore the point that we have all heard[,] and that you and I have heard directly from our nation's voting rights enforcement agents in the Voting Section at DoJ, that accurate data is critical especially for small areas. The data must reflect what is seen in the real world because it is used to change how the real world interacts with itself and with its government. This does not mean I do not understand our obligation to protect the public's data, it just appears that in our zeal to protect the data we are harming the very same people we are protecting.

> —Ex. 1, September 30, 2020 Email from James Whitehorne, Chief of the Redistricting & Voting Rights Data Office at the U.S. Census Bureau, to John Abowd, Associate Director and Chief Scientist at the U.S. Census Bureau

"Harming the very same people we are protecting" is an apt description of the Census Bureau's use of differential privacy to intentionally inject error into the census data States need to redistrict. In preparation for redistricting, the Secretary of Commerce and Alabama agreed on "a plan identifying the geographic areas for which specific tabulations of population [within Alabama] are desired." 13 U.S.C. § 141(c). And the Secretary's duty to provide "tabulations of population" for each census block in the State requires her to provide the actual numbers of people Defendants counted during the census—not some other numbers generated by the Bureau. Don't just take our word for it. As late as September 2020, the Bureau stated that it was "*legally mandated*" to provide "*error-free disclosure of block-level population totals under Public Law 94*."<sup>1</sup>

What changed? Dr. John Abowd—the Bureau's Chief Scientist and chief proponent of differential privacy—determined that if accurate block-level population totals were reported to the

<sup>&</sup>lt;sup>1</sup> See Ex. 2, U.S. Census Bureau, Statistical Expertise & General Research Topics 11 (Sept. 2020), https://perma.cc/Z6JK-RLY5 (emphasis added); see also Ex. 3, Simson L. Garfinkel, U.S. Census Bureau, Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the 2018 Endto-End Test (Continued) (Sept. 15, 2017), https://perma.cc/H325-562D (referencing 2000 DOJ agreement to hold block population, block voting age population, and block householders and vacancies counties exact for State redistricting tabulations).

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 11 of 65

States, his preferred method of disclosure avoidance would "break."<sup>2</sup> The solution? Break the law instead. As Abowd put it, no more "garbage about being legally required to hold certain populations invariant."<sup>3</sup>

Having disposed of that "garbage," Defendants now present a radical new reading of Section 141(c) by which States are no longer entitled to accurate tabulations of population—only tabulations that "represent the sub-state population" in some way, shape, or form that suits Defendants' fancy. Resp., Doc. 41 at 32.<sup>4</sup> When Plaintiffs asked in their motion for a preliminary injunction whether "assigning every Alabamian to Birmingham would violate Alabama's right under subsection 141(c) to tabulations of population for specific geographic areas," doc. 3 at 44, the question was meant to be rhetorical. They didn't realize this was actually Defendants' position.

The absurdities don't stop there. Defendants claim that Alabama isn't harmed by the Bureau's delay or its inaccurate tabulations because the State could always conduct its own census. Resp., Doc. 41 at 35-36. (Elsewhere they provide important qualifiers: The Department of Justice will use the *Census Bureau*'s data "for enforcement of the Voting Rights Act," and conducting a census costs billions of dollars. *Id.* at 5, 22. No harm indeed.) They assert that the Bureau's decision to blow past the March 31 deadline for delivering the tabulations is not final agency action, *id.* at 69—even though the Bureau did in fact blow past the March 31 deadline. And they assert that the final tabulations will be free from "improper or partisan manipulation," Resp., doc. 41 at 15, even as they boast that the accuracy of certain racial or ethnic groups will take priority over the accuracy of other groups when it comes to the zero-sum allocation of the privacy loss budget, Abowd Decl., doc. 41-1 at 35. But Congress required accurate tabulations drawn from the actual

<sup>&</sup>lt;sup>2</sup> See Ex. 4, Email from John Abowd to Ron Jarmin (Jul. 7, 2020).

<sup>&</sup>lt;sup>3</sup> See Ex. 5, Email from John Abowd to Gary Benedetto (Jul. 28, 2020).

<sup>&</sup>lt;sup>4</sup> Citations are to the ECF-stamped page numbers.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 12 of 65

enumeration precisely to guard against the dangers of such manipulation. *See generally* 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). Indeed, the Census Bureau's "tuning" of the top-down algorithm through "large numbers of TDA runs"<sup>5</sup> calls to mind Justice Kagan's concern about the appearance of manipulation in the redistricting context: "While bygone mapmakers may have drafted three or four alternative districting plans, today's mapmakers can generate thousands of possibilities at the touch of a key—and then choose the one giving their party maximum advantage...." *Rucho v. Common Cause*, 139 S. Ct. 2484, 2513 (2019) (Kagan, J., dissenting).

It almost seems like that's the entire point, because the Bureau's stated reason for implementing differential privacy is so spectacularly flawed. As Dr. Steven Ruggles, the developer of the world's largest census database and director of the Institute for Social Research and Data Innovation at the University of Minnesota (and a former member of the Census Scientific Advisory Committee) shows, the problem differential privacy was allegedly meant to fix does not exist. *See* Ex. 6, Ruggles Expert Report at 5-11. For instance, most "matches" in the Bureau's database reconstruction simulation appear to have occurred purely by chance, or at similar rates. *Id.* at 7-8. In fact, Ruggles concludes, "the 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." *Id.* at 8. As for the bottom-line number—the Bureau's purported reidentification of 17% of census respondents—it seems a bit ridiculous to conclude that guessing a person's race or ethnicity correctly 1 out of 6 times constitutes "re-identification" in any real sense

<sup>&</sup>lt;sup>5</sup> U.S. Census Bureau, *Meeting Redistricting Data Requirements: Accuracy Targets* (Apr. 7, 2021), https://perma.cc/Y3LP-NVR7.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 13 of 65

of the word. Were it otherwise, the Census Bureau has been flagrantly violating its privacy obligations for nearly a century. Moreover, even that 17% number is too generous, for as Defendant Ron Jarmin noted in 2019, the outside attacker couldn't even be sure she was right that often without "access to confidential internal Census Bureau information."<sup>6</sup>

On the other hand, the results of unnecessarily skewing population numbers and other characteristics will harm the Plaintiffs. To be sure, the Bureau asserts that as it races to complete its novel system (months behind schedule), it will likely address many of the anomalies that have appeared in its previous four sample datasets, in part by increasing its privacy loss budget to make the final tabulations less inaccurate. But the final tabulations will still be intentionally skewed, and there will still be no way for Alabama to tell how many people live where based on the actual enumeration. That will cause significant problems when it comes to using the flawed data to "change how the real world interacts with itself and with its government" (as Whitehorne put it),<sup>7</sup> such as drawing Congressional and legislative districts that comply with one-person, one-vote and the Voting Rights Act. And by increasing the privacy loss budget, the final privacy risk (as Defendants define it) under differential privacy will likely be similar to-or worse than-the privacy risk (again, as Defendants define it) present when only traditional disclosure avoidance methods are used. See Ex. 6, Ruggles Expert Report at 16-19. In other words, because the final tabulations will still be intentionally and unlawfully skewed, these changes by the Bureau will simply mean that Defendants are now sure to violate *both* their accuracy obligation under 13 U.S.C. § 141(c) and their privacy obligations under 13 U.S.C. §§ 8-9 (as Defendants misinterpret them) to protect every single characteristic against an all-knowing theoretical adversary. Worse still, unlike with

<sup>&</sup>lt;sup>6</sup> Ex. 7, Ron Jarmin, U.S. Census Bureau, *Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census* (Feb. 15, 2019), https://perma.cc/54R7-YJQS.

<sup>&</sup>lt;sup>7</sup> Ex. 1, Email from James Whitehorne to John Abowd (Sept. 30, 2020).

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 14 of 65

past methods of disclosure avoidance, differential privacy's added losses in privacy "may be trivial or they may be very risky under real world conditions" because "Differential Privacy does not distinguish" between the two. *See* Amicus Br. of Prof. Jane Bambauer, Doc. 33 at 17.

Defendants need not violate either statute. Contra Defendants' blindered reading, the privacy protections of 13 U.S.C. §§ 8 and 9 do not mandate the protection of every characteristic from the made-up threat of an attacker who knows every bit of information except for one. Not only would such a reading produce absurd results and make the Bureau's past 92 years of releases patently illegal, but it forces the Census Act into conflict with itself, simultaneously requiring and forbidding Defendants from providing accurate redistricting data to the States. Congress did no such thing. Defendants can comply with both statutory requirements, just as they have done in decades past.

As for remedy, Plaintiffs do not ask the Court to take over the decennial census. This is no programmatic attack. Rather, Plaintiffs ask for injunctive relief that stems the harm caused by two discrete actions by Defendants: delaying release of the redistricting data by five months, and altering the reported population numbers through the application of differential privacy. As discussed below, there are many ways to remedy these problems. So long as they comply with the law, *how* Defendants clean up their mess is their choice. But the Court can provide relief by ordering Defendants to hold population counts invariant at the block level and to deliver the population tabulations no later than July 31, 2021—within the three-month window contemplated by Congress between the release of the apportionment data (now April 30) and the release of the redistricting tabulations. Plaintiffs will be irreparably harmed if the Court does not do so.

#### **ARGUMENT**

### I. Plaintiffs Have Standing To Sue.

"The 'irreducible constitutional minimum' of standing consists of three elements." *Spokeo, Inc. v. Robins*, 136 S. Ct. 1540, 1547 (2016) (quoting *Lujan v. Defs. of Wildlife*, 504 U.S. 555, 560 (1992)). Plaintiffs must have "(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." *Id.* And because "standing is not dispensed in gross," at least one Plaintiff "must demonstrate standing for each claim he seeks to press." *Town of Chester v. Laroe Ests., Inc.*, 137 S. Ct. 1645, 1650 (2017) (citations omitted). Plaintiffs have met these requirements for both their differential privacy and their delay claims.

## A. Defendants' Application of Differential Privacy Harms Plaintiffs. This Court Can Redress These Harms.

In their Complaint and preliminary injunction motion, Plaintiffs demonstrated a number of harms flowing from the Bureau's implementation of differential privacy. *See* Compl., Doc. 1 at 32-37; Mot., Doc. 3 at 40-47, 61-67. Those harms can be abated by this Court. Defendants none-theless contend that Plaintiffs haven't suffered a single injury, that any injury Plaintiffs have experienced was not caused by Defendants, and that the Court is powerless to do anything even if Defendants are responsible for Plaintiffs' harm. Resp., Doc. 41 at 24-46. No, no, and no.

### 1. Differential Privacy Harms Plaintiffs.

"To establish injury in fact, a plaintiff must show that he or she suffered 'an invasion of a legally protected interest' that is 'concrete and particularized' and 'actual or imminent, not conjectural or hypothetical." *Spokeo*, 136 S. Ct. at 1548 (quoting *Lujan*, 504 U.S. at 560). Plaintiffs have done so here. Among other things, the Bureau's application of differential privacy: breaches the Secretary's commitment to provide Alabama with accurate tabulations of population; causes

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 16 of 65

the State to rely to its detriment on the Secretary's promise to deliver population tabulations; harms Plaintiffs' Congressionally granted informational right to receive accurate tabulations; impedes the State's sovereign interest in drawing fair and lawful legislative and congressional districts; subjects the State to near-certain litigation; unfairly distributes federal funding within and among States; creates a substantial risk that the individual Plaintiffs' voting rights will be diluted; and harms the individual Plaintiffs' right to accurate redistricting data under Section 209.

Informational Harms and Breach of Commitment. In the Census Act, Congress created a multi-year process by which States may obtain "specific tabulations of population" for the "geographic areas" desired by the State to redistrict. See 13 U.S.C. § 141(c). The process begins "not later than April 1 of the fourth year preceding the decennial census date," when the Secretary establishes criteria for State plans. Id. It continues "not later than 3 years before the decennial census date," when the State officers "having initial responsibility for the legislative apportionment or districting of each State" submit their plans to the Secretary. Id. And it concludes "within one year after the decennial census date," when the Secretary must transmit the redistricting tabulations "to each respective State." Id. Congress thus provided the participating States with an informational right to population tabulations with accurate population counts that can be used for redistricting. And in Section 209, Congress extended that right to "any resident of a State whose congressional representation or district could be changed as a result of the use of a statistical method." Pub. L. No. 105-119, § 209(d)(1)-(2). The violation of this right is an Article III harm. See Spokeo, 136 S. Ct. 1549-50 (citing FEC v. Akins, 524 U.S. 11, 20-25 (1998) (confirming that a group of voters' "inability to obtain information" that Congress had decided to make public is a sufficient injury in fact to satisfy Article III) and Pub. Citizen v. Dep't of Just., 491 U.S. 440, 449 (1989) (holding that advocacy organizations' inability to obtain information subject to disclosure under

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 17 of 65

the Federal Advisory Committee Act "constitutes a sufficiently distinct injury to provide standing to sue")). Indeed, "the illegal deprivation of information" is "[a] prime example" of "the violation of a procedural right set out in a statute [that] will necessarily result in the harm that Congress was trying to prevent." *Muransky v. Godiva Chocolatier, Inc.*, 979 F.3d 917, 929-30 (11th Cir. 2020) (en banc). It "is not a 'bare' procedural violation, or one that is 'divorced from any concrete harm." *Id.* at 930. Because Sections 141(c) and 209 "protect[] against a lack of information," Defendants' "denial of access to" accurate redistricting data "is a concrete injury" to all Plaintiffs. *Id.* 

In addition, Alabama has long relied on accurate population tabulations from the Secretary to draw its legislative and congressional districts. *See* Mot., Doc. 3 at 15-20. It did so again for this upcoming redistricting cycle. *See* Loftin Decl., Doc. 3-1 at 2-3; McClendon Decl., Doc. 3-2 at 2-3. The State submitted "a plan identifying the geographic areas for which specific tabulations of population are desired," 13 U.S.C. § 141(c), and relied on the Secretary to provide the tabulations as required by Congress. Apparently, that reliance was to the State's detriment. Now Defendants claim that Plaintiffs will not be harmed—and the Secretary's statutory duty will be fulfilled—so long as the Secretary produces some kind of data in tabular form. *See* Resp., Doc. 41 at 24. Then they resort to argument-by-gotcha and reason that because Plaintiffs referred to the inaccurate tabulations as "tabulations," Plaintiffs have "themselves prove[d] the point" that the Secretary will fulfill her statutory mandate if she produces tabulations with any data in them. *See id.* at 24 ("Plaintiffs do not dispute that the Secretary will provide to the State data in such an arranged form [as a tabulation]. Hence, Alabama will receive 'tabulations."); *id.* at 25 ("[Plaintiffs] admit that the tabulations that the Secretary will deliver are, in fact, 'tabulations of population.").

This reasoning proves a bit too much. By reading out the word "population" in "tabulation[s] of population," Defendants seem to contend that the Census Bureau really could "assign[]

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 18 of 65

every Alabamian to Birmingham" in the population tabulations. *See* Mot., Doc. 3 at 44. That not only seems wrong, it *is* wrong for at least two reasons. First, it's clear from the text of the statute that, at the very least, the tabulations of *population* must include the *population* counts as enumerated. Congress did not require the Secretary to provide worthless tabulations "to the officers or public bodies having responsibility for legislative apportionment or districting" in each State as an empty exercise in paper shuffling. 13 U.S.C. § 141(c); *cf.* Antonin Scalia & Bryan A. Garner, *Reading Law: The Interpretation of Legal Texts* 63-64 (2012) (explaining that "[a] textually permissible interpretation that furthers rather than obstructs the document's purpose should be favored").<sup>8</sup> Indeed, not to borrow too much from Defendants' flawed reasoning, but it seems notable that even the Census Bureau refers to the tabulations produced under 13 U.S.C. § 141(c) as "P.L. 94-171 Redistricting Data."<sup>9</sup> The data are *for* redistricting. The Census Bureau cannot simply produce bad data in tabular form and tell the States "you're welcome."

Second, Alabama has long relied on the tabulations to include the population counts as enumerated, unaffected by any disclosure avoidance methods. *See* Ex. 6, Ruggles Expert Report at 2-3. Indeed, not too long ago the Census Bureau itself acknowledged that the total population counts and the block voting age population counts *had* to remain invariant to comply with 13 U.S.C. § 141(c). As Simson L. Garfinkel, the Census Bureau's Chief of the Center for Disclosure Avoidance Research, noted, that had long been the understanding of both the Census Bureau and the Department of Justice:

<sup>&</sup>lt;sup>8</sup> For the same reason, Defendants' rejoinder regarding sampling of the P.L. 94-171 data fails. Resp., Doc. 41 at 28-29. As 13 U.S.C. § 141 makes clear, the population tabulations used for redistricting are tied to the decennial census and must reflect the "population" there enumerated. Thus, just as the Bureau cannot fudge the numbers it gives to the President for the apportionment, neither can it fudge the numbers it gives to the States for redistricting. That is true with or without § 195's additional prohibition of sampling in the apportionment data.

<sup>&</sup>lt;sup>9</sup> See, e.g., U.S. Census Bureau, Decennial Census P.L. 94-171 Redistricting Data (Mar. 15, 2021), https://perma.cc/MF5A-CMQK.

# Some queries must be privacy preserving. Some queries must be exact ("invariant")



Ex. 3, 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 (Continued)* (Sept. 15, 2017), https://perma.cc/H325-562D.

Notably, Garfinkel gave his presentation on September 15, 2017—a year *after* Alabama submitted its request for specific population tabulations. *See* Loftin Decl., Doc. 3-1 at 3. And as Garfinkel notes, the Bureau's understanding that it was required to hold block population and block voting age population invariant was no new thing; it was memorialized in an agreement with the Department of Justice in 2000. Accordingly, because the Secretary had reported true population numbers in the P.L. 94-171 data for so long, Plaintiffs reasonably relied on the Secretary doing so again. After all, just months ago, the Bureau was still stating that it was "*legally mandated*" to provide "*error-free disclosure of block-level population totals under Public Law 94*."<sup>10</sup>

Yet in November 2020—four years after Alabama submitted its request, and just four months before the redistricting data were due—the Bureau abruptly changed course and determined that it would *not* report the enumerated population and voting age population counts to the

<sup>&</sup>lt;sup>10</sup> See Ex. 2, U.S. Census Bureau, Statistical Expertise & General Research Topics 11 (Sept. 2020), https://perma.cc/Z6JK-RLY5 (emphasis added).

States. *See* Ex. 8, U.S. Census Bureau, Invariants Set for 2020 Census Data Products (Nov. 24, 2020), https://perma.cc/JET7-WDCL. Why? Not because the law changed, but because—according to Abowd—holding the population counts invariant would "break" differential privacy. *See* Ex. 4, Email from John Abowd to Ron Jarmin (Jul. 7, 2020). Thus, it appears Defendants decided to break the law instead. The Bureau prioritized Abowd's project and quietly removed its past references to the law's requirements. *Compare, e.g.*, Aref N. Dajani et al., Presentation to Census Scientific Advisory Committee, *The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau* 6 (Sept. 2017), https://perma.cc/5GKQ-LVRR (noting that, "[b]y agreement with the Department of Justice (2000), the Census Bureau will provide exact counts at the Census block level for the following variables: Number of people: total, age 18+ (voting age), and less than age 18...."), with John M. Abowd et al., *The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau* (July 2020), https://perma.cc/634N-4XPU (removing reference to these legal requirements and the DOJ agreement).

These actions were unlawful because they ignored Plaintiffs' cognizable reliance interests for no good reason (in fact, for very bad reasons). *See Dep't of Homeland Sec. v. Regents of the Univ. of Cal.*, 140 S. Ct. 1891, 1914 (2020) (noting that "[w]hen an agency changes course ... it must be cognizant that longstanding policies may have engendered serious reliance interests that must be taken into account" (cleaned up and citation omitted)). But the Bureau's decisions also harmed Plaintiffs. Again, the State reasonably thought it would be getting the "legally mandated error-free disclosure of block-level population totals" that Defendants had previously provided.<sup>11</sup> Then, after the enumeration had been completed and the redistricting data were months away from being released, the Bureau abruptly changed course. As a result, Alabama will either be forced to

<sup>&</sup>lt;sup>11</sup> See Ex. 2, U.S. Census Bureau, Statistical Expertise & General Research Topics 11 (Sept. 2020), https://perma.cc/Z6JK-RLY5 (emphasis added).

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 21 of 65

redistrict with tabulations it and everyone else knows are false, or (as Defendants suggest) conduct its own census. These are significant harms caused by Defendants' actions.

Defendants also assert that Plaintiffs will not be harmed by the false population tabulations because the statistical error introduced by differential privacy may in fact cause the numbers to reflect more closely "the actual (but inherently unknowable) population of a given census block on Census Day." Resp., Doc. 41 at 31. Of course, it's just as likely that the introduced error will cause the reported numbers to deviate even more significantly from "the actual (but inherently unknowable) population." Regardless, for redistricting it's the "actual enumeration" that matters. Speaking at a time when the intra-state population counts were held constant, the Supreme Court explained: "Even if one cannot say with certainty that one district is larger than another merely because it has a higher census count, one *can* say with certainty that the district with a larger census count is more likely to be larger than the other district than it is to be smaller or the same size." Karcher v. Daggett, 462 U.S. 725, 738 (1983). Hence, "because the census count represents the best population data available, it is the only basis for good-faith attempts to achieve population equality." Id. Once the numbers in the population tabulations are divorced from the actual enumeration, though, all bets are off. Either States will be forced to conduct their own censuses if they desire an actual count, or they can redistrict with numbers that do not reflect the census count. Either way they will be harmed.

*Sovereign injury*. Alabama has a sovereign interest in drawing fair congressional and legislative districts that comply with one-person, one-vote and the requirements of the Voting Rights Act. *See* Mot., Doc. 3 at 44-46. Defendants discount this interest by pointing out that the State hasn't been sued—yet—and suggesting that if Alabama "believes that the future census redistricting data will be unsuitable for apportionment and redistricting, Alabama may conduct its own census." Resp., Doc. 41 at 34-35. Neither argument is valid. The State's interest in drawing fair districts for its residents is independent of whether it will get sued—though there *is* a substantial risk that it will get sued. *E.g., Ala. Legis. Black Caucus v. Alabama*, 989 F. Supp. 2d 1227, 1235 (M.D. Ala. 2013) (collecting cases), *vacated and remanded on other grounds*, 575 U.S. 254 (2015). Just as "a legislative body has a judicially cognizable interest in matters affecting its composition," *U.S. House of Representatives v. U.S. Dep't of Com.*, 11 F. Supp. 2d 76, 87 (D.D.C. 1998), so too does a State have an interest in how its electoral districts are composed. Blinding the State to where its residents live harms that interest.

As for Defendants' "build your own census" argument, the State participates in the P.L. 94-171 process so it doesn't have to conduct its own census. P.L. 94-171 was enacted in 1975 precisely so States wouldn't have to conduct their own censuses, but could instead "cooperate with the Census Bureau in defining the geographic areas for which they would receive small area population counts for redistricting." Margo Anderson, *Baker v. Carr, the Census, and the Political and Statistical Geography of the United States: The Origin and Impact of Public Law 94-171*, 62 Case W. Res. L. Rev. 1153, 1168 (2012). That's the Congressional bargain on which Alabama has long relied. In any event, Defendants indicate that conducting a census is time-consuming and expensive work that cost the Bureau over \$15 billion. *See* Thieme Decl., Doc. 41-2 at 3. Even if the State could conduct its own census at this late hour (it can't), Defendants are not picking up the tab, so either way the State is harmed.

*Federal Funding.* Inaccurate census data will also affect the distribution of federal funding by sending those funds to the wrong place. *See* Mot., Doc. 3 at 63-66. Defendants' rejoinder is to say that "to the extent that Alabama's funding would be affected by differential privacy, it will result in a *windfall* to the State" because rural and urban areas will be treated differently by the

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 23 of 65

differential privacy algorithm. Resp., Doc. 41 at 37. This misunderstands Plaintiffs' harm. As Representative Aderholt explained, "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." Aderholt Decl., Doc. 3-11 at 5. The misallocation of federal funding harms the State—which will have both winners *and* losers—as well as Representative Aderholt in his work in Congress. *Id*.

As for Defendants' invocation of the Supreme Court's decision in Trump v. New York, 141 S. Ct. 530 (2020), two significant differences make reliance on the decision inapposite. For one, the Court's decision that it lacked jurisdiction to hear the plaintiffs' challenge to the President's memorandum excluding illegal aliens from the apportionment base relied on both standing doctrine and ripeness concerns. See id. at 535. "Ripeness reflects constitutional considerations that implicate Article III limitations on judicial power, as well as prudential reasons for refusing to exercise jurisdiction." Stolt-Nielsen S.A. v. AnimalFeeds Int'l Corp., 559 U.S. 662, 670 n.2 (2010) (quotations omitted). To the extent the Court's opinion rested on prudential ripeness, such concerns have no purchase here because Section 209 "eliminated any prudential concerns in this case." See Dep't of Com. v. U.S. House of Representatives, 525 U.S. 316, 328 (1999). For another, the challenged action here is far more certain than the action in Trump. In that case, the "President qualified his directive by providing that the Secretary should gather information 'to the extent practicable," and it was entirely unclear whether or how the President would exclude illegal aliens from the apportionment base. Trump, 141 S. Ct. at 535. But here the die has been cast. The Bureau has committed to using differential privacy and to holding only three categories of data constanttotal population of each State, total housing units at the census block level, and the number of group quarters facilities by type at the census block level. Because the intrastate population counts

will *not* be reported as enumerated, they—or at least a significant portion of them—will necessarily be false, making it substantially likely that federal resources will be misallocated within the State.

*Vote Dilution.* It is also substantially likely that the voting rights of the individual Plaintiffs will be harmed. *See* Mot., Doc. 3 at 46-47. Defendants dismiss these concerns because the individual Plaintiffs' votes haven't yet been diluted since their districts haven't been drawn. They also argue that Plaintiffs "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." Resp., Doc. 41 at 52.

The Supreme Court's decision in *Department of Commerce v. U.S. House of Representatives*, 525 U.S. 316 (1999), resolves both objections. That case (which was actually two cases combined on appeal) concerned a pre-census challenge to the Bureau's plan to use statistical sampling to determine the population for the apportionment. *Id.* at 320. The Court held that the Bureau's plan was unlawful. *Id.* at 334. Before reaching that question, though, the Court specifically found that the plaintiffs in the underlying *Gavin* case had demonstrated standing "on the basis of the expected effects of the use of sampling in the 2000 census *on intrastate redistricting.*" *Id.* at 332 (emphasis added). Citing *Karcher*'s statement that the census count provides "the only basis for good-faith attempts to achieve population equality," *id.* at 334 (quoting *Karcher*, 462 U.S. at 438), the Court held that the parties living in the districts likely to be affected by sampling "have a strong claim that they will be injured by the Bureau's plan because their votes will be diluted vis-à-vis residents of counties with larger 'undercount' rates," *id.* The Court also explained: "[T]he *threat* of vote dilution … is concrete and actual or imminent, not conjectural or hypothetical." *Id.* at 332 (cleaned up, citation omitted, and emphasis added)). So it is here. Plaintiffs have shown that the application of differential privacy—like the Bureau's previous plan to introduce sampling—is substantially likely to harm the individual Plaintiffs because of its expected effects on intrastate redistricting. *See* Mot., Doc. 3 at 33-34. That threat is not conjectural or hypothetical; it is real because, even if the Bureau improves its differential privacy algorithm, the Bureau does not plan to hold the intrastate population counts invariant.<sup>12</sup> And while none of the three individual Plaintiffs knows for sure whether Bureau officials will select redistricting data that favors or harms him (though Mr. Green, as an urban voter, will be particularly likely to be harmed, *cf.* Resp., doc. 41 at 37-38), Plaintiffs need not wait for the roulette pistol to fire before seeking relief. *See Davis v. FEC*, 554 U.S. 724, 735 (2008) ("[T]he injury required for standing need not be actualized."). And, particularly since "we will never be able to assess the ... accuracy of the" manipulated data "by comparing it to the results of" the actual count, *House of Reps.*, 525 U.S. at 349 (Scalia, J., concurring in part), it is "not necessary for this Court to wait until the [districts have been drawn] to consider the issues presented here, because such a pause would result in extreme—possibly irremediable—hardship." *Id.* at 332.

## 2. Defendants Are Applying Differential Privacy, Thus Causing Plaintiffs' Harm.

Because the Census Bureau is the one implementing differential privacy to skew the population tabulations, it is clear Plaintiffs' harms are "fairly traceable to the challenged conduct" of Defendants. *Spokeo*, 136 S. Ct. at 1547. "[E]ven harms that flow indirectly from the action in

<sup>&</sup>lt;sup>12</sup> Indeed, Abowd indicates in his declaration that the *fifth* iteration of the differential privacy experiment (to be released after the date of this filing) will make certain changes. Abowd now says that "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." Doc. 41-1 at 37 ¶70. And he specifies that this information will be "tuned" to be accurate only for the largest racial group. While it is unclear what the term "voting districts" here refers to (Congressional District, state legislative district, state senate district, voting precinct, etc.), there are many "spine" and "off-spine" geographies that have fewer than 500 people, and there is no indication of accuracy for the second-largest racial group in any "voting district." As Professor Andy Beveridge has noted, fully one out of every five "voter tabulation districts" in the United States have fewer than 500 people, nearly one out of every three census places have fewer than 500 people, and 99% of census blocks have fewer than 500 people. Andy Beveridge, *Sixteen States Sue to Block Census Bureau Data Privacy Method* (Apr. 19, 2021), https://perma.cc/WEJ9-TEXY.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 26 of 65

question can be said to be fairly traceable to that action for standing purposes." *Focus on the Family v. Pinellas Suncoast Transit Auth.*, 344 F.3d 1263, 1273 (11th Cir. 2003) (quotation marks and citation omitted).

Defendants' counterarguments are unavailing. The first is that Plaintiffs—not Defendants—shot themselves in the foot because the Alabama Legislature erroneously relied on the Census Bureau to produce accurate population tabulations instead of ordering a census itself. Resp., Doc. 41 at 43. This is silly. The reliance was reasonable, long-standing, and supported by statute and Supreme Court precedent. Telling the State to go build its own multi-billion-dollar census in response to Defendants' last-minute breach hardly severs the line of traceability.

Defendants' second argument is that "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." Resp., Doc. 41 at 43. But it's not Plaintiffs' job to tell Defendants *how* to clean up their mess. And as detailed below, the Bureau has many options to produce accurate tabulations in compliance with § 141(c), Plaintiffs' reliance interests, and the Census Act's privacy requirements. For instance, the Bureau could revert to the disclosure avoidance methods used in 2010 with great success. *See* Mot., Doc. 3 at 20-23; Ex. 6, Ruggles Expert Report at 13-16; *see also* Amicus Br. of Prof. Jane Bambauer, Doc. 33 at 17-23. Or it could alter the amount of data, or the privacy settings, for releases of the other more granular data that gives the Bureau concern. *See* Ex. 6, Ruggles Expert Report at 14-16. Or Defendants could fix the "problem" some other way. The important part is that these examples show that Plaintiffs' harms are caused by Defendants and that, as explained next, the harms are redressable by this Court.

# 3. The Harms Caused By Differential Privacy Are Redressable.

This Court can redress the harms caused by Defendants' application of differential privacy. "To have Article III standing, a plaintiff need not demonstrate anything 'more than ... a substantial likelihood' of redressability." *Wilding v. DNC Servs. Corp.*, 941 F.3d 1116, 1126727 (11th Cir. 2019) (alterations in original) (quoting *Duke Power Co. v. Envtl. Study Grp.*, 438 U.S. 59, 79 (1978)). The remedy here is to order the "legally mandated error-free disclosure of block-level population totals under Public Law 94."<sup>13</sup> That would redress Plaintiffs' harms.

Defendants claim they can't do this because they are required by 13 U.S.C. §§ 8 and 9 to skew the numbers. *See* Resp., Doc. 41 at 45-46. As detailed below, that's neither factually nor legally correct. It's not factually correct because the re-identification attacks the Census Bureau relies on did not show a true threat to privacy. *See* Ex. 6, Ruggles Expert Report at 5-11. And it's not legally correct because Defendants fundamentally misunderstand Congress's privacy directive. As the Census Bureau itself explained in *Baldridge v. Shapiro*, 455 U.S. 345 (1982), "all *raw* census data relating to particular individuals or establishments (as distinguished from *aggregate statistical data*) is subject to the confidentiality mandate of the Census Act." Reply Br. of Petrs., *Baldridge v. Shapiro*, No. 80-1436, 1981 WL 389926, at \*3 (Nov. 25, 1981) (emphasis added). The population tabulations Plaintiffs seek are not "raw census data," but "aggregate statistical data."

Moreover, even if Defendants were right in their interpretation of the Census Act's privacy requirements, that does not mean that they can violate the Act's requirement in § 141(c) to give accurate redistricting population tabulations to the States. The Bureau must comply with both mandates. Again, one way to do that is this: present the P.L. 94-171 redistricting data unperturbed, or

<sup>&</sup>lt;sup>13</sup> See Ex. 2, U.S. Census Bureau, Statistical Expertise & General Research Topics 11 (Sept. 2020), https://perma.cc/Z6JK-RLY5 (emphasis added).

at least with the population and voting-age population counts held invariant. *See* Ex. 6, Ruggles Expert Report at 15-16. Even by the Bureau's standards, that data alone would not allow anyone to succeed in a re-identification attack because they contain so few variables. *Id.* Rather, to succeed in the attack, an adversary would also need the tables the Bureau publishes in *other* datasets, such as Summary File 1 and Summary File 2, which the Bureau typically publishes in the months and years following release of the redistricting data.<sup>14</sup> It is *those* datasets that contain the vast majority of characteristics that, when combined with outside information, present the theoretical privacy risk claimed by Defendants. *Id.* As Abowd noted in his declaration, the Census Bureau used the 2010 Summary File 1 dataset—not simply the P.L. 94-171 redistricting tabulations—to conduct its re-identification attack simulation. That dataset included "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." Abowd

<sup>&</sup>lt;sup>14</sup> The Census Bureau describes the Summary File 1 dataset this way: "Summary File 1 (SF 1) contains the data compiled from the questions asked of all people and about every housing unit. Population items include sex, age, race, Hispanic or Latino origin, household relationship, household type, household size, family type, family size, and group quarters. Housing items include occupancy status, vacancy status, and tenure (whether a housing unit is owner-occupied or renter-occupied). There are 177 population tables (identified with a 'P') and 58 housing tables (identified with an 'H') shown down to the block level; 82 population tables (identified with a 'PCT') and 4 housing tables (identified with an 'HCT') shown down to the census tract level; and 10 population tables (identified with a 'PCO') shown down to the county level, for a total of 331 tables. The SF 1 Urban/Rural Update added 2 PCT tables, increasing the total number to 333 tables. There are 14 population tables and 4 housing tables shown down to the block level and 5 population tables shown down to the census tract level that are repeated by the major race and Hispanic or Latino groups. SF 1 includes population and housing characteristics for the total population, population totals for an extensive list of race (American Indian and Alaska Native tribes, Asian, and Native Hawaiian and Other Pacific Islander) and Hispanic or Latino groups, and population and housing characteristics for a limited list of race and Hispanic or Latino groups. Population and housing items may be cross-tabulated." U.S. Census Bureau, Summary File 1 Dataset, https://perma.cc/38JG-RAFD.

The Summary File 2 dataset has additional detail, though provided only at the census tract level, rather than down to the census block level: "Summary File 2 (SF 2) contains the 100-percent data (the information compiled from the questions asked of all people and about every housing unit). Population items include sex, age, race, Hispanic or Latino, household relationship, and group quarters. Housing items include occupancy status, vacancy status, and tenure (owner occupied or renter occupied). SF 2 includes population characteristics, such as sex by age, average household size, house-hold type, relationship by household type (including living alone), unmarried-partner households, nonrelatives by household type, and own children under 18 years by family type and age. The file includes housing characteristics, such as tenure, tenure by age of householder, and tenure by household size for occupied housing units." U.S. Census Bureau, Summary File 2 Dataset, https://perma.cc/S276-KY59.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 29 of 65

Decl., Doc. 41-1 at 63 (emphasis added). Simply put, "[t]he sparse data available in the redistricting file is insufficient to allow database reconstruction; only in combination with other tables can database reconstruction even be attempted." Ex. 6, Ruggles Expert Report at 15.

Notably, unlike with the redistricting tabulations, Defendants are under no obligation under P.L. 94-171 to report the SF1 and SF2 data in any specific form. And to the extent other federal laws require any of the more granular detail in those datasets, the Bureau has the time and discretion to ensure the data are reported responsibly.<sup>15</sup> This means that the Bureau does not get a free pass to skew the required P.L. 94-171 tabulations simply because it chooses to release "an enormous number of [additional] statistics calculated from its collected data." Abowd Decl., Doc. 41-1 at 8. If the Bureau chooses to apply differential privacy to those other datasets that will not impact redistricting—fine. Or if the Bureau chooses to release fewer statistics in the other datasets to reduce the risk of privacy disclosure—also fine. The Bureau retains tremendous discretion regarding those choices, and there is ample time for the Bureau to consider the privacy protections needed for the other datasets after it releases the P.L. 94-171 tabulations. *See* Ex. 6, Ruggles Expert Report at 15-16.<sup>16</sup> But what the Bureau cannot do is what it plans to do: skew the population counts in the redistricting data Congress requires Defendants to provide the States. An order preventing Defendants from doing that would give Plaintiffs relief.

Defendants object that "[a]n order enjoining the use of differential privacy would also only extend the Bureau's delay in providing redistricting data." Resp., Doc. 41 at 46. But that assumes

<sup>&</sup>lt;sup>15</sup> There may be other federal statutes that require certain of the information provided in SF1 and SF2 to be released, and the Bureau regularly engages with federal agencies, state governments, and other data users to determine what information is legally required and in what format. The SF1 and SF2 files are historically released months after the redistricting data and are generally not used in redistricting. *See* Ex. 6, Ruggles Expert Report at 15-16.

<sup>&</sup>lt;sup>16</sup> Abowd even indicated his desire to hold back these releases "until there is consensus that the underlying data are fit-for-use." Ex. 14, Email from John Abowd to Danah Boyd (Oct. 7, 2020).

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 30 of 65

that the alternative disclosure avoidance methods will need to be much altered so they can inject as much error into the data as differential privacy does. As explained below, that's incorrect.

Nor have Defendants satisfactorily explained why reverting to the 2010 disclosure avoidance methods would cause additional delay. Abowd first notes that the Bureau would need "to conduct the requisite software development and testing" before using the disclosure avoidance software previously tested and successfully used in 2010. Doc. 41-1 at 46. And he states that "[t]he 2020 Census's system architecture is completely different from that used in the 2010 Census, and it is thus not possible to simply 'plug in' the disclosure-avoidance system used in 2010." *Id.* But he gives no estimate as to how long this would take or why, and there's good reason to think the 2010 methods could be applied more quickly than still-in-development differential privacy.

First, as Ruggles notes, "[t]he main differences in the system architecture between 2010 and 2020 pertain to the disclosure control software itself," meaning that the Census Edited File (CEF) "will be functionally the same as the 2010 CEF." Ex. 6, Ruggles Expert Report at 13. "Accordingly, there should be few barriers to applying the 2010 software to the 2020 data" because the disclosure avoidance method is applied to the CEF to create what is called the Hundred-percent Detail File (in 2010) or the Microdata Detail File (in 2020). *Id.* "In terms of structure," he explains, the two files are equivalent, "so it is unlikely that using traditional statistical disclosure controls would slow down the tabulation phase." *Id.* Defendants' claimed delay thus appears to be "highly exaggerated." *Id.* 

Second, before Ohio sued Defendants on February 25, 2021, over their plan to delay delivery of redistricting data, Defendants' position was that they could not deliver until September 30. *See Ohio v. Raimondo*, No. 3:21-cv-064, 2021 WL 1118049 (S.D. Ohio Mar. 24, 2021). But by the time their response to Ohio's preliminary injunction motion was due, their sense of what was

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 31 of 65

possible had expanded. Turns out "[i]t would be possible to have fully reviewed redistricting data for all States available by mid to late August," though "the data would be in an older format of data the Census Bureau developed decades ago." *See* Whitehorne Decl., Doc. 11-2 at 9, *Ohio*, No. 3:21-cv-064 (S.D. Ohio Mar. 12, 2021). This litigation too may help Defendants realize they can "jump higher, run faster, or lift more" to timely implement lawful disclosure avoidance methods. Resp., Doc. 41 at 48. And Defendants' ability to use "an older format of data" seems to suggest that the 2010 disclosure avoidance methods can be utilized once again as well.

Finally, enjoining differential privacy need not cause delay because the P.L. 94-171 data is not the cause of Defendants' purported privacy dilemma. The reason the Bureau claims to be unable to release the P.L. 94-171 tabulations quickly is because it needs to skew the data; and it claims it needs to do that because of the privacy risks posed by forthcoming SF1 and SF2 datasets. So, again, here's a solution: release the redistricting tabulations unperturbed or after applying simple disclosure avoidance techniques, and then adjust the SF1 and SF2 datasets as needed. That would allow the Bureau to provide the required redistricting tabulations quickly and accurately. The Court need not impose that exact solution, of course. But it does show that the Defendants can remedy Plaintiffs' harm if the Court requires them to produce population tabulations with true population counts. How Defendants achieve that result is up to them.

# B. Defendants' Delay in Releasing the Tabulations of Populations Harms Plaintiffs. This Court Can Redress That Harm.

Plaintiffs have also shown Article III standing to challenge Defendants' decision to unlawfully delay delivering the redistricting tabulations beyond the statutory deadline. Plaintiffs have been harmed by that decision, and the magnitude of the harm *grows* with each passing day Plaintiffs are not afforded relief. Because this Court can provide such relief, it has jurisdiction to do so.

#### 1. The Delay Harms Plaintiffs.

*First and foremost*, the Bureau's delay prevents the State from effecting its redistricting process in accordance with the Alabama Constitution. See Mot., Doc. 3 at 17-20. Because the Legislature cannot redistrict until the P.L. 94-171 redistricting tabulations are released, Defendants' delay is preventing the Legislature from meeting its constitutional obligations. *See* Loftin Decl., Doc. 3-1. The Bureau has thus prevented the State from "effectuating" its law, thus causing "irreparable injury." *Maryland v. King*, 567 U.S. 1301, 1303 (2012) (Roberts, C.J., in chambers); *see also Muransky*, 979 F.3d at 930-31 (noting that "allegations of wasted time can state a concrete harm for standing purposes").

Defendants point out that the Alabama Constitution contains a failsafe switch to "provide for enumeration of inhabitants for purpose of apportionment of representatives and senators." Ala. Const. art. IX, § 201. But it does so only if the decennial census is either "not taken," or, "when taken," the census "be not full and satisfactory" "as to this state." *Id.* Neither condition applies. There is no dispute that the Bureau is in fact "tak[ing]" a "decennial census of the United States," so the first condition is inapplicable. As to the second, the decennial census has not yet been "taken"—indeed, the Bureau's contention that the census remains incomplete implies as much meaning the second condition fares no better. Regardless, Defendants' suggestion that Alabama build its own census to evade the harm caused by the Bureau is not a serious one. Forcing the State to invest untold amounts of time and money to create its own redistricting numbers—and then be sued for failing to use the "gold standard" of census data—amounts to a "concrete and particularized" injury, too. *Spokeo*, 136 S. Ct. at 1548.

Second, the Bureau's actions compromise "the State's interest in protecting the integrity and reliability of the electoral process." *Crawford v. Marion Cty. Election Bd.*, 553 U.S. 181, 191 (2008) (lead op. of Stevens, J.). Specifically, delivering redistricting data on September 30 will 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. As Clay Helms, the Director of Elections for the Alabama Secretary of State's Office, explained, that is unlikely to be enough time because the reassignments typically take up to six months. See Mot., Doc. 3 at 67; Helms Decl., Doc. 3-3 at 2-5. Thus, he stated, the delay "could result in one or more 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." Doc. 3-3 at 3-4. Defendants take Helms to task for using the word "could," and assert that the threatened injury is speculative. Resp., Doc. 41 at 41. Lest there be any doubt, Helms has now clarified that he "did not mean to suggest these eventualities are not likely." Ex. 9, Suppl. Helms Decl. ¶¶ 1-2. Rather, he "believe[s] that at least one of these eventualities is very likely to occur-and that any combination of them could occur-if the Bureau forces the Boards of Registrars to reassign voters on the compressed timeline the Bureau has suggested." Id.; see also id. at ¶ 4 (explaining why the Board of Registrars cannot begin the reassignment process now).

*Finally*, as Representative Aderholt explained in his declaration, *see* Aderholt Decl., doc. 3-11, the Bureau's delay harms him both as a candidate and a voter. Defendants retort that the reduced campaign time should redound to his benefit him as an incumbent, Resp., doc. 41 at 37-38, but that is both speculative and irrelevant: The delay forces Rep. Aderholt to adopt new, second-best alternatives to his campaign strategy, thereby injuring him. *See, e.g., Zimmerman v. City of Austin*, 881 F.3d 378, 390 (5th Cir. 2018) ("[C]hanging one's campaign plans or strategies in response to an allegedly injurious law can itself be a sufficient injury to confer standing.").

### 2. The Delay Is Traceable to the Bureau's February 12 Decision.

These harms are "fairly traceable to the challenged conduct of the defendant[s]," *Spokeo*, 136 S. Ct. at 1547, because Plaintiffs' delay-related harms flow from the Bureau's failure to deliver redistricting data on time. *Focus on the Family*, 344 F.3d at 1273. While Defendants insist that Plaintiffs cannot show traceability "because redistricting is ultimately the responsibility of the State," Resp., doc. 41 at 44, if that were true, the plaintiffs in *House of Representatives* would have never "established standing on the basis of the expected effects of the use of sampling in the 2000 census on intrastate redistricting." 525 U.S. at 332. As for Defendants' contention that traceability is severed "because [Plaintiffs] identify no feasible alternative to producing redistricting data by September 30, 2021," Resp., doc. 41 at 44, while the viability of alternative arrangements might implicate *redressability*, they do not refute traceability. Plaintiffs' injuries are caused by the delayed delivery of redistricting data, and that delay is caused by Defendants.

### *3. The Delay Is Redressable.*

It is substantially likely that Plaintiffs' delay harms can be redressed by this Court. See Wilding, 941 F.3d at 1126–27. The Bureau deliberately violated § 141(c)'s statutory deadline, and that decision immediately harmed Plaintiffs. Plaintiffs' injuries are continuing, so the sooner Plaintiffs receive the relief they seek the less harm they will ultimately suffer. And conversely, the longer the Bureau protracts its statutory breach, the more harm Plaintiffs incur. See, e.g., Aderholt Decl., Doc. 3-11 ¶ 22 ("The Census Bureau's delays have a cascading effect on my reelection."); see also Ex. 9, Suppl. Helms Decl. ¶¶ 1-7. And, as explained above, the Bureau can provide at least partial relief.

Defendants respond that "apportionment will be entirely in Congress's hands"—not the Bureau's—"to accept or reject." Resp., Doc. 41 at 47 (citing 2 U.S.C. § 2a(b)). To the extent De-

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 35 of 65

fendants are arguing that Plaintiffs' claims are not redressable because ultimate authority for congressional apportionment lies beyond the Bureau, the Supreme Court has twice addressed and expressly rejected similar arguments. *See Utah v. Evans*, 536 U.S. 452, 464 (2002); *Franklin v. Massachusetts*, 505 U.S. 788, 803 (1992). And Plaintiffs need not wait on Congress before drawing state house or senate districts, so that delay is solely Defendants' doing. Besides, Congress—like Plaintiffs—is waiting for the Bureau to provide the data. Thus, the more delay the Bureau imposes on Congress, the more delay Plaintiffs suffer in turn.

More fundamentally, Defendants allege that Plaintiffs lack standing because "Alabama seeks the impossible." Resp., Doc. 41 at 48. To support the assertion, Defendants claim that "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." *Id.* at 46-47. But Plaintiffs' requested relief is not so limited. Rather, Plaintiffs asked the Court to "enjoin Defendants from delaying the release of data beyond *the earliest possible date this Court determines equitable* and that will allow the State to use the redistricting data during the redistricting process." Compl., Doc. 1 at 52 (emphasis added). It is not too late to ameliorate Plaintiffs' harms. And the relief Plaintiffs seek is not "impossible" in any event. Resp., Doc. 41 at 48. We know the Bureau can produce redistricting data within three months of providing apportionment data because it has always done so. And there is every reason to believe that it could do so again if it dropped differential privacy.<sup>17</sup> As Ruggles explains in his declaration, "[g]iven that the most complex remaining aspect of census processing that remains to be completed is the

<sup>&</sup>lt;sup>17</sup> Indeed, Michael Thieme's declaration indicates that the CUF was completed March 10, and that the CEF will be completed by mid-June. *See* Doc. 41-2 at 17-20. The CUF file contains the accurate population information by census block and is used to produce the statewide population totals for the apportionment count. The CEF then takes this file and appends detailed demographic data. In 2010, the Bureau produced the P.L. 94-171 redistricting data from the CEF in 27 days. Doc. 41-2 at 21. If a similar schedule is kept this year, the P.L. 94-171 data could be disseminated by July 31, 2021.

final execution of differential privacy, it is more plausible that substituting the simpler, well-understood protocols from 2010 could actually speed the processing time." Ex. 6, Ruggles Expert Report at 14. The Bureau has several viable disclosure-avoidance methods aside from differential privacy. *Id.* at 13-16. It should choose one, and in so doing mitigate the harm of its delay.

## II. Plaintiffs Are Likely To Prevail On The Merits.

Plaintiffs are entitled to a preliminary injunction because they have demonstrated that they are likely to succeed on the merits, that they are likely to suffer irreparable harm unless this Court prevents it, that the injury to Plaintiffs outweighs the inconvenience to Defendants caused by an injunction, and that the public interest favors injunctive relief. See Siegel v. Lepore, 234 F.3d 1163, 1176 (11th Cir. 2000) (en banc). Defendants contend that Plaintiffs really seek a permanent injunction, Resp., doc. 41 at 48, but that's neither relevant nor true. It's not relevant because the standards are "essentially the same," with "the exception that the plaintiff must show ... actual success" on the merits for a permanent injunction-which Plaintiffs have done for all but their arbitrary-and-capricious claims, for which Defendants have not provided the administrative record. Amoco Prod. Co. v. Vill. of Gambell, 480 U.S. 531, 546 n.12 (1987). And it's not true because it is Defendants who have upended the "the last uncontested status quo between the parties," Canal Auth. of Fla. v. Callaway, 489 F.2d 567, 576 (5th Cir. 1974), by (1) announcing just months before the redistricting tabulations were due that it would skew the population counts at the census block level, and (2) deciding that it would not meet the March 31 release deadline and would instead publish the final tabulations in September. Those decisions were illegal and must be enjoined.

### A. Defendants' Application of Differential Privacy Is Unlawful.

### 1. The Census Act Requires Accurate Tabulations of Population.

There are two main points of disagreement between Plaintiffs and Defendants when it comes to the Census Act's requirements. The first is what "tabulation[s] of population" means in
§ 141(c). *Compare* Mot., Doc. 3 at 40-44, *with* Resp., Doc. 41 at 24-31. Defendants say it means data that bear some relation—however scant—to the enumerated population numbers arranged in tabular form. *See* Resp., Doc. 41 at 32. As discussed above, Plaintiffs contend that Congress likely did not pass such a useless law and that, instead, the tabulations must reflect the "actual enumeration"—including, most importantly, the invariant intrastate population counts. And a few months ago, when differential privacy proved incapable of delivering these numbers, the Bureau too agreed that Section 141(c) imposed a "legal[] mandate[]" of "error-free disclosure of block-level population totals."<sup>18</sup> The law didn't change in the interim. Only Defendants' adoption of differential privacy did.

The second disagreement concerns the meaning of the Census Act's privacy requirements codified at 13 U.S.C. §§ 8 and 9. The relevant provisions of those sections provide:

[T]he Secretary may furnish copies of tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent, and may make special statistical compilations and surveys, for departments, agencies, and establishments of the Federal Government, the government of the District of Columbia, the government of any possession or area (including political subdivisions thereof) referred to in section 191(a) of this title, State or local agencies, or other public and private persons and agencies, upon payment of the actual or estimated cost of such work....

13 U.S.C. § 8(b).

Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison, 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.A. § 9(a).

<sup>&</sup>lt;sup>18</sup> See Ex. 2, U.S. Census Bureau, Statistical Expertise & General Research Topics 11 (Sept. 2020), https://perma.cc/Z6JK-RLY5; *see also* Ex. 3, Garfinkel presentation (noting that "per 2000 Department of Justice letter to the Director" of the Bureau, block population and block voting age population "must be exact").

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 38 of 65

Defendants interpret these requirements to mean that the Census Bureau can never publish characteristic data that matches a characteristic of a respondent. Abowd provides a helpful example: "Any block where the voting-age data are either all 'yes' or all 'no'"—that is, where either everyone is or is not of voting age—"is an exact attribute disclosure assignable to all persons living in that block on [census day]" because it reveals an individual characteristic: voting age. <sup>19</sup> In other words, by publishing data that the hundred people on the census block are all of voting age, we have now learned that the hundred people on the census block are all of voting age. Even more fundamentally, we can add, by publishing data that a hundred people live on the census block, we have now learned that those 100 people live on the census block. According to Defendants, such tautologies violate the Census Act. *See* Ex. 6, Ruggles Expert Report at 12.

If that sounds absurd, that's because it is. In fact, it's unclear clear why Defendants' interpretation would allow them to publish any invariant data at all. The Secretary plans to release the actual population count of how many people live in Alabama. But wouldn't that also reveal an individual characteristic regarding each resident—the state of residence? "Under this new interpretation, the Census Bureau has been in flagrant violation of the law ever since 1929," when Congress enacted the predecessor to §§ 8 and 9. *Id*. That is because "[e]very tabulation of the characteristics of the population necessarily reveals the attributes of individuals," and "[e]very census from 1790 to 2010 has published attributes based on exact numbers counted in the census." *Id*. And again, the longstanding understanding by Defendants has been that such releases are not only okay, but required. *See* Ex. 3, Garfinkel presentation. The only plausible explanation for De-

<sup>&</sup>lt;sup>19</sup> See Ex. 10, John Abowd, Tweetorial: Reconstruction-abetted re-identification attacks and other traditional vulner-abilities (Apr. 7, 2019), https://perma.cc/LUH3-QC8R

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 39 of 65

fendants' abrupt change is that the statutory requirement ran headlong into the buzzsaw of differential privacy. As Abowd put it, holding the population counts invariant would "break"<sup>20</sup> differential privacy, so the "garbage about being legally required to hold certain populations invariant"<sup>21</sup> had to go.

"That history matters." *Dep't of Com. v. New York*, 139 S. Ct. 2551, 2567 (2019). The "open, widespread, and unchallenged" practice sheds light on the provision's meaning. *See id.* (quoting *NLRB v. Noel Canning*, 573 U.S. 513, 572 (2014) (Scalia, J., concurring in judgment)); *see also Wisconsin v. City of New York*, 517 U.S. 1, 21 (1996) (noting "importance of historical practice" in census context). The Court should thus reject Defendants' "invitation to measure the" legality of the census "by a standard that would seem to render every census since" 1929 unlawful. *New York*, 139 S. Ct. at 2567. Whatever protections §§ 8 and 9 provide, the provisions cannot mean what Defendants say they mean—that the Bureau is forbidden from releasing accurate population and voting age population counts at the census block level.

Other clues bolster this conclusion. For one, Defendants' "sterile literalism ... loses sight of the forest for the trees." *Pictet Overseas Inc. v. Helvetia Trust*, 905 F.3d 1183, 1191 (11th Cir. 2018) (W. Pryor, J., concurring) (quoting *N.Y. Tr. Co v. Comm'r of Internal Revenue*, 68 F.2d 19, 20 (2d Cir. 1933) (Hand, J.)). The forest is the rest of the Census Act, the context in which §§ 8 and 9 appear. *See* Scalia & Garner, *Reading Law* 167 ("Perhaps no interpretive fault is more common than the failure to follow the whole-text canon, which calls on the judicial interpreter to consider the entire text, in view of its structure and of the physical and logical relation of its many parts"). Because § 141(c) requires the Secretary to provide tabulations of population reflecting the actual enumeration—at the very least, how many people live where within the State—it is doubtful

<sup>&</sup>lt;sup>20</sup> See Ex. 4, Email from John Abowd to Ron Jarmin (Jul. 7, 2020).

<sup>&</sup>lt;sup>21</sup> See Ex. 5, Email from John Abowd to Gary Benedetto (Jul. 28, 2020).

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 40 of 65

that another provision of the same Act forbids the Secretary from fulfilling that duty. *See also* Scalia & Garner, *Reading Law* 63-64 ("A textually permissible interpretation that furthers rather than obstructs the document's purpose should be favored"). The same is also true of the Secretary's duty to provide accurate census data under other statutes. *E.g.*, 7 U.S.C. § 361c(c)(2) (allocating funding based on census data); *id.* § 2663(b)(4) (same); 49 U.S.C. § 5305(d)(1)(A)(i) (same); *id.* § 5311(c)(3)(B)(iii) (same). There is no reason to force those provisions into artificial conflict.

Defendants rely heavily on the Supreme Court's decision in *Baldridge v. Shapiro*, 455 U.S. 345 (1982), to support their alternative reading. Resp., Doc. 41 at 61-62. But that case simply confirms that Defendants have seriously misread the statute. In *Baldridge*, city officials sought the master address list the Census Bureau used to conduct the 1980 decennial census—the "addresses, householders' names, number of housing units, types of census inquiry, and, where applicable, the vacancy status of the unit." *Id.* at 349. Understandably, the Court determined that such information was protected by §§ 8 and 9. *Id.* at 359. It explained that "*raw* data reported by or on behalf of individuals was to be held confidential and not available for disclosure." *Id.* at 355 (emphasis added); *see also id.* at 358 (noting that the history "of the Census Act reveals a congressional intent to protect the confidentiality of census information by prohibiting disclosure of *raw* census data reported by or on behalf of individuals" (emphasis added)).

As the Census Bureau's brief in *Baldridge* explained, the key word in those sentences is "raw": "The government's position is that all raw census data relating to particular individuals or establishments (*as distinguished from aggregate statistical data*) is subject to the confidentiality mandate of the Census Act." Reply Br. of Petrs., *Baldridge v. Shapiro*, No. 80-1436, 1981 WL 389926, at \*3 (Nov. 25, 1981) (emphasis added). The Bureau then defined "aggregate statistical materials" to mean "totals and subtotals of numerical figures." *Id.* at \*7. That is precisely what

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 41 of 65

Plaintiffs seek: the totals and subtotals of numerical figures reflecting the actual population counts within the State. The plaintiffs in *Baldridge* got that. *See* Br. of Petrs, *Baldridge v. Shapiro*, No. 80-1436, 1981 WL 389922, at \*42 (Jul. 13, 1981) ("[T]he Census Bureau provided each local jurisdiction ... with aggregate information for each enumeration district, including the number of housing units, the number of vacant units and population totals."). So should these Plaintiffs.

There is no reason Defendants cannot provide that data again. To return to Abowd's votingage population example, publishing a table showing that everyone in a census block is of voting age does not reveal "raw data" for two reasons. First, the data is processed. Raw data would publish an individual's actual age or birthday—whatever the respondent reported to the Census. That's not what the Bureau publishes. Instead, what the tabulation shows is the processed data stating whether the respondents were of voting age. That's allowed by the Census Act. Second, the data is aggregated—again, not raw. Bob's voting age characteristic is aggregated with Jim's and so on down the census block. It's the aggregated statistic that is published. Again, that's not raw data. It can be published.

To be sure, the Census Bureau understandably takes additional measures to protect the identity of census respondents. Under characteristic swapping, for instance, the Bureau targets individuals whose unique characteristics make them susceptible to being identified, and it swaps the identifying characteristics with those of a respondent in a different census block. *See* Mot., Doc. 3 at 21-22; *see also* Ex. 6, Ruggles Expert Report at 17 (providing example of couple who lived on Liberty Island whose age and race characteristics were swapped). Under any reasonable interpretation of §§ 8 and 9, and certainly under any real-world identification concern, the Bureau's traditional disclosure avoidance methods have protected identities well.

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 42 of 65

The Bureau responds that its studies show otherwise, and points to a simulated reidentification attack against the 2010 census data that "precisely reconstructed approximately 46% of the 308,745,538 records with their exact race, ethnicity, sex, and age." Resp., Doc. 41 at 11. But this study suffers from basic analytical flaws, which is likely why it has never been subjected to the rigors of peer review. Now that Abowd's declaration has revealed more about the simulated attack, it is even clearer that the attack failed in numerous ways.

First, as Abowd states in his declaration, 57% of the population in the 2010 census were unique as to block, race, ethnicity, sex, and age. Doc. 41-1 at 65-66. That means that 43% of the population shared a census block with someone of the same race, ethnicity, sex, and age. Thus, if the Bureau had randomly selected 1000 respondents from the 2010 census data, the Bureau would expect to match them with another census respondent sharing all their characteristics a full *43%* of the time. This means that the Bureau's "reconstruction" attack—the whole reason for differential privacy—produced a "match rate" of 46% that was hardly any better than random chance.

Ruggles confirmed this with a simple simulation.<sup>22</sup> Looking only at age-sex combinations (for which 44% of the population have unique characteristics for their census block, Abowd Decl., Doc. 41-1 at 65-66), Ruggles found "randomly chosen age-sex combinations would match someone on any given block 54.9% of the time." Ex. 6, Ruggles Expert Report at 7. According to Ruggles, "[t]his means that[] the Census Bureau would have been 'correct' on age and sex 55% of the time if they had never looked at the tabular data from 2010, and had instead just assigned ages and sexes to their hypothetical population at random." *Id.* Even more stunning, when Ruggles

<sup>&</sup>lt;sup>22</sup> "To estimate the percentage of random age-sex combinations that would match someone on a block by chance, [Ruggles] generated 1000 simulated blocks and populated them with random draws from the 2010 single-year-of-age and sex distribution. The simulated blocks conformed to the population-weighted size distribution of blocks observed in the 2010 census. [He] then randomly drew 1000 new age-sex combinations and searched for each of them in each of the 1000 simulated blocks. In 54.9% of cases [he found someone in the simulated block who exactly matched the random age-sex combination. The simulation source code and supporting data files are available at <a href="http://use-ers.hist.umn.edu/~ruggles/censim.html">http://use-ers.hist.umn.edu/~ruggles/censim.html</a>." Ex. 6, Ruggles Expert Report at 7 n.4.

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 43 of 65

"assign[ed] everyone on each block the most frequent race and ethnicity of the block, and then randomly [drew] age-sex combinations, then 42.7% of cases in the hypothetical database would be expected to match on *all four characteristics* to a respondent on the same block." *Id.* at 7-8 (emphasis added). Compare that to the Bureau's match rate in its reidentification attack: 46%. As Ruggles concludes, "despite the Census Bureau's enormous investment of resources and computing power, the 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." *Id.* at 8.

Second, even accepting that the Bureau had "reconstructed" any useful data, attempts to match it with commercially available data were largely unsuccessful. While the Bureau claims to have "re-identified" 16.85% of the population in its re-identification attack simulation, "[t]hat claim is irresponsible." Ex. 6, Ruggles Expert Report at 9. The "reconstructed" data were usually incorrect and usually failed to "match even the block, age and sex of anyone identified in outside commercial sources." Id. at 10. And "[i]n the minority of cases where a hypothetical reconstructed individual d[id] match the block, age, and sex of someone in the commercial data, it usually turn[ed] out that the person identified in the commercial data was not actually enumerated on that block in the census." *Id.* In other words, if an attacker is wrong 5 times out of 6 in guessing the remaining characteristics-race and ethnicity-of census respondents, that level of uncertainty cannot constitute "re-identification" in any real sense of the word. See also Amicus Br. of Prof. Jane Bambauer, Doc. 33 at 19 ("The Census Bureau's most recent examination of the 2010 census records found greater number of apparent matches, but the attempted attacks were similarly lousy in making accurate matches.... [T]he researchers were able to make matches on 45% of the records.... However, the vast majority of those matches (62%) were wrong." (citation omitted)).

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 44 of 65

Again, were it otherwise, the Census Bureau has been flagrantly violating its privacy obligations for nearly a century.

Moreover, as now Acting Director of the Census Bureau Ron Jarmin noted in 2019, the outside attacker couldn't even be sure he was right 1 in 6 times without "access to confidential internal Census Bureau information."23 No one outside the Bureau has that information-and if they did, they'd have no reason to run a reconstruction attack—so this is an important admission. Ruggles breaks it down this way: "Reidentification means confirming the identity of a particular individual and revealing their characteristics without reference to non-public internal census files. It would be impossible to positively identify the characteristics of any particular individual using the database reconstruction without access to non-public internal census information. Accordingly, the Census Bureau's database reconstruction experiment demonstrates that reidentification based on the published census tables is not feasible." Ex. 6, Ruggles Expert Report at 9. JASON-the third-party group whom Defendants use to bolster the supposed need for differential privacy, see Resp., Doc. 41 at 11-12—made a similar point: "Such reconstruction of the microdata [from the simulation attack] is not yet a violation of Title 13 since no personal data (e.g. names, addresses, etc.) are used when these tables are built." JASON, Formal Privacy Methods for the 2020 Census 29 (Apr. 2020), https://perma.cc/G8ZM-YNN6. Thus, the re-identification attack actually demonstrated the *efficacy* of the Bureau's 2010 disclosure avoidance methods.

Finally, the failure of the reconstruction attack is shown by considering that even under a preposterous worst-case scenario, there is no plausible harm and no reason for conducting the attack. Abowd explains that the "harm" of the re-identification attack is that the "attacker can learn self-response race and ethnicity"— assuming the attacker already knows a respondent's name, age,

<sup>&</sup>lt;sup>23</sup> Ex. 7, Ron Jarmin, U.S. Census Bureau, *Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census* (Feb. 15, 2019), https://perma.cc/54R7-YJQS.

sex, and address. Doc. 41-1 at 107. But the attacker is successful at making this guess only 17% of the time. And even in a counterfactual "worst-case scenario" in which the attacker has access to perfect 2010 census data on name, address, sex, and age, the attacker correctly guesses race and ethnicity only 58% of the time. *Id.* at 71 ¶24. This uncertainty shows that no one has been "re-identified." It also shows why no one would ever attempt such an attack, for if the attacker who already possessed perfect data on name, address, sex, and age wanted to make a good guess on race and ethnicity, she could do much better by simply assuming that the respondent possesses the same race and ethnicity as the majority of people in her census block. Ex. 6, Ruggles Expert Report at 7-8. For example, if the attacker knew that respondents lived in a census block where 75% of people were white, he could correctly guess their race 75% of the time by guessing white. He'd have no reason to reconstruct a database that would lead to *less* accurate guesses. Thus, the threat that justified differential privacy and skewed population counts was a wholly phantom menace.

It's also worth pointing out that differential privacy does not even allow the Census Bureau to meet its own interpretation of §§ 8 and 9. Unlike with characteristic swapping, which intentionally targets unique identifiers that make a response a high-risk for personal identification, differential privacy is "a blunt and inefficient instrument for disclosure control." *Id.* at 13. The effect is that as accuracy increases, privacy—as defined by Defendants—decreases. But as the demonstration data show, to have even remotely usable data, the privacy loss budget must be ratcheted up dramatically. Defendants boast they're doing exactly that, *see* Resp., Doc. 41 at 62, but they never explain how the corresponding losses in privacy can cohere with their understanding of §§ 8 or 9 if those provisions prohibit the release of *any* individual characteristic.

The Bureau's (purported) re-identification attacks prove the point. The Census Bureau seems to be planning a privacy loss budget of 12.3 or less for the P.L. 94 redistricting data file.

See Ex. 6, Ruggles Expert Report at 16. And the new demonstration product is set to have a budget of 10.3. See Abowd Decl., Doc. 41-1 at 28. But these epsilon levels are "far higher than ... ordinarily contemplated by privacy researchers," who generally suggest "runs from 0.01 to 5.0." Ex. 6, Ruggles Expert Report at 16. "Accordingly, one would expect that  $\varepsilon$ =12.3 would provide a relatively low level of data security." Id. That seems to be true: The Census Bureau re-ran the database reconstruction against the 2010 census data after applying differential privacy and concluded that "approximately 7.5% of the noise-infused population would have 'confirmed re-identification' using the same methodology as was employed in the original Census Bureau database reconstruction experiment." Id. at 17. That rate is both (1) unlawful under Defendants' interpretation, and (2) in the same vicinity as the rate of purported re-identifications under past methods. *Id.*; cf. Amicus Br. of Prof. Jane Bambauer, Doc. 33 at 2 ("[W]hen 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.)"). Given that, Defendants have not shown why they must violate § 141(c) just so they can also violate §§ 8 and 9 (as they understand them). JASON warned the Bureau in March 2020 that this would happen, but Defendants plowed ahead anyway. See JASON, supra at 94 ("The proposed use of DP in the 2020 Census, which is by now almost certain, will bring the mandates of Section 214 [which incorporates § 9] and Section 141 into conflict to a substantially greater degree than previously.... 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 policy—should it prove to be controversial—lies elsewhere.").

It gets worse, though. Real-world privacy protection is in fact likely to be *worse* off under differential privacy. As privacy expert Professor Jane Bambauer explains in her amicus brief, when

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 47 of 65

it comes to actual (as opposed to theoretical) threats, "Differential Privacy is *inferior* to the traditional methods that model different threat scenarios and quantify risks under a range of assumptions." Doc. 33 at 16 (emphasis added). "[T]he quantitative precision of Differential Privacy is actually a drawback," she says, because it "has the patina of mathematical elegance without actually quantifying privacy risks of the sort that most people care about." *Id.* at 16. Under differential privacy, for instance, the greater privacy losses caused by a more robust privacy loss budget "may be trivial or they may be very risky under real world conditions," but it's impossible to know because "Differential Privacy does not distinguish between these two." *Id.* at 17. Importantly, that is not true for traditional methods of disclosure avoidance. *Id.* 

Ruggles makes the same point: "By contrast [with traditional disclosure avoidance methods], differential privacy makes no distinctions between high-risk and low-risk cases, so it infuses noise equally across characteristics and populations. This means that to achieve a given level of disclosure control, differential privacy must introduce far more error than would be needed using traditional statistical disclosure control." Ex. 6, Ruggles Expert Report at 18. And, he notes, "[i]t has long been recognized ... that there is no direct relationship between the level of  $\varepsilon$  [or the privacy-loss budget] and the risk of disclosing identities." *Id.* at 16; *id.* ("Because differential privacy does not target variables and circumstances that are vulnerable to attack, in some datasets with strong differential privacy (low  $\varepsilon$ ), disclosure control can be weak."). For this reason, JASON suggested that the Census Bureau "apply swapping as performed for the 2010 census" "[i]n addition to the use of Differential Privacy ... so that no unexpected weakness of Differential Privacy as applied can result in a 2020 census with less protection tha[n] 2010." JASON, *supra*, at 98. Apparently, the Bureau rejected this advice. Otherwise, it could not claim that reverting to 2010 methods would somehow cause any delay. *Cf.* Abowd Decl., Doc. 41-1 at 46.

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 48 of 65

Lastly, differential privacy also presents a significant risk of political manipulation in the census results. Under differential privacy, Census workers have wide discretion to "manage the accuracy" (as Abowd puts it) of the census data as they see fit. *See* Ex. 10, Abowd at 1. Thus, by "tuning" the algorithm just so, or allotting the privacy loss budget in particular ways, bureaucrats can ensure that the population counts and other data are reported more accurately for certain groups—and reported *less* accurately for others. Think rural voters have too much say? There's a technical fix for that. Don't like older voters, or members of a certain racial group? Ditto. And the best (or worst) part is: No one will ever know of the manipulation because the accurate population numbers will never be released.

Such fears could be easy to discount if the Census Bureau *wasn't already prioritizing the accuracy of the data for certain racial and ethnic groups over others*. To be sure, nothing has come to light so far of such blatantly nefarious attempts as the examples above. But manipulation can happen with the best of intentions. For example, after hearing for months and months of the (valid) concerns from the American Indian and Alaska Native communities concerning the impact of differential privacy, Census officials determined that tribal areas would be "given sufficient privacy-loss budget to ensure that those populations are presented accurate to the number of persons in the unit column ... essentially invariant and the same precision as the state populations themselves." Abowd Decl., Doc. 41-1 at 34. Plaintiffs agree with part of that decision: The Bureau should provide the AIAN population counts as enumerated. The problem is that allocation of the privacy loss budget is zero-sum. Abowd confirms this: "[T]his solution still requires balancing accuracy and privacy-loss overall. All characteristics cannot have large privacy-loss budget allocations at every geographic level." *Id.* at 34-35. Thus making the data more accurate for one ethnic or racial group will necessarily make the data less accurate for other ethnic or racial groups. As

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 49 of 65

one Census official noted in opposing the plan to prioritize the AIAN communities: "We cannot promise to do something that blatantly gives one racial group an advantage at the expense of all others." *See* Ex. 11, Email from Benjamin Overholt to Nathaniel Cogley (Oct. 22, 2020); *cf. Shaw v. Reno*, 509 U.S. 630, 643-44 (1993) ("A racial classification, regardless of purported motivation, is presumptively invalid and can be upheld only upon an extraordinary justification.").

Alas, it appears those objections were overruled. They shouldn't have been. To give Defendants the "power … to select among various estimation techniques having credible (or even incredible) 'expert' support is to give the[m] … the power to distort" census results in ways that favor certain groups or parties over others. *U.S. House of Representatives*, 525 U.S. at 348-49 (Scalia, J., concurring in part). A reading of the Census Act that requires such manipulation cannot be the best one, particularly in light of the long history of Congress acting to *prevent* such manipulation. *See* Pub. L. No. 105-119, § 209(a)(7) (recognizing 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"); *cf. Utah*, 536 U.S. at 497-506 (Thomas, J., concurring in part and dissenting in part) (detailing the history of census manipulation and the Founders' efforts to combat it).

For all these reasons, the Court should reject Defendants' misreading of the Census Act. Subsection 141(c) requires the Secretary to provide States with tabulations of the population that include the actual population counts, not intentionally distorted approximations of those numbers. And the privacy guarantees in §§ 8 and 9 do not forbid such release, and thus do not require the Census Bureau's adoption of differential privacy.<sup>24</sup>

<sup>&</sup>lt;sup>24</sup> The individual Plaintiffs' Equal Protection claims are also likely to succeed for the reasons stated in Plaintiffs' initial motion. *See* Doc. 3 at 46-47. Defendants raise mainly standing objections in response, Doc. 41 at 52, which are addressed above.

# 2. Section 209 Provides Plaintiffs With a Cause of Action.

As explained in Plaintiffs' motion for a preliminary injunction, *see* Mot., Doc. 3 at 48-49, in § 209 of the Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act of 1998, Congress created a cause of action for persons harmed by the Census Bureau's use of unlawful statistical methods to seek relief in Court. That cause of action reads:

Any person aggrieved by the use of any statistical method in violation of the Constitution or any provision of law (other than this Act), 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.

Pub. L. No. 105-119, § 209(b), 111 Stat. 2440 (codified at 13 U.S.C. § 141 note).

Defendants do not dispute that Rep. Aderholt, Mr. Green, and Mr. Williams each has a cause of action under this provision to challenge differential privacy. *See* Resp., Doc. 41 at 49-50.<sup>25</sup> Defendants argue that Alabama lacks a similar cause of action, but even if true, the individual Plaintiffs have a cause of action under § 209, and the State has a cause of action under the APA (as do Plaintiffs). Each claim is thus properly before the Court. *See* Compl, Doc. 1 at 45-49.

Defendants' objection lacks merit in any event. Congress enacted the cause of action in § 209 because it recognized that "it would be impracticable *for the States* to obtain, and the court of the United States to provide, meaningful relief after" the census was complete. Pub. L. No. 105-119, § 209(a)(8) (emphasis added). It thus allowed "[a]ny person aggrieved by the use of any statistical method" to file suit, *id.* § 209(b), and expanded the concept of "person aggrieved" to "*include*[]" "(1) any resident of a State whose congressional representation or district could be

<sup>&</sup>lt;sup>25</sup> In their opposition to Plaintiffs' request for a three-judge court, Defendants did contest that the application of differential privacy qualifies as a "statistical method" within the meaning of § 209. *See* Opp., Doc. 23. The Court preliminarily overruled those objections when it convened the three-judge panel and determined that Plaintiffs "sufficiently alleged an action covered by § 209 and 28 U.S.C. § 2284." Order, Doc. 27 at 6. For the reasons stated in Plaintiffs' reply in support of the three-judge court request, *see* doc. 25, Defendants' objection was based on a factually incorrect understanding of "statistical method."

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 51 of 65

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," *id.* § 209(d) (emphasis added). Considering that § 209 was created for the benefit of the States, and then appended as a note to § 141(c)—which was also created for the benefit the States, *see* Anderson, *supra*—the context of the Act confirms that Congress's inclusion of other aggrieved persons did not exclude the parties most likely to be affected by the Bureau's malfeasance: the States. That context provides sufficient "statutory intent" to rebut the generic presumption that States are not statutory "persons." *Return Mail, Inc. v. U.S. Postal Serv.*, 139 S. Ct. 1853, 1862 (2019); *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).

# 3. Plaintiffs Brought Their Claims Neither Too Early Nor Too Late.

The Court should also set aside Defendants' decision to apply differential privacy to the redistricting data as unlawful and arbitrary and capricious. *See* 5 U.S.C. § 706(2)(A), (B), (C). The merits of these claims, and Defendants' objections to them, are sufficiently addressed above and in the initial motion. *See* Mot., Doc. 3 at 50-55. Here, a few notes in response to Defendants' too-early, too-late procedural trap (for these claims and others) will suffice.

*First*, Defendants contend that Plaintiffs do not challenge agency "action" under the APA, but rather have launched a programmatic attack. Resp., Doc. 41 at 54-55. As the Supreme Court has reminded, "[t]he bite in the phrase 'final action' ... is not in the word 'action,' which is meant to cover comprehensively every manner in which an agency may exercise its power." *Whitman v. Am. Trucking Assocs.*, 531 U.S. 457, 478 (2001) (citations omitted). And the APA defines agency

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 52 of 65

action broadly to include "the whole or part of an agency rule, order, license, sanction relief, or the equivalent or denial thereof, or failure to act." 5 U.S.C. § 551(13). A "rule," in turn, includes "the whole or a part of an agency statement of general or particular applicability and future effect designed to implement, interpret or prescribe law or policy or describing the organization, procedure, or practice requirements of an agency." *Id.* § 551(4). These definitions encompass the Bureau's discrete decision to apply differential privacy in a manner that skews the intrastate population and voting-age population counts at the district block level.

As for Defendants' "programmatic attack" characterization, a programmatic attack occurs where, for example, a litigant "seek[s] *wholesale* improvement of [a] program by court decree," *Lujan*, 497 U.S. at 891, or asks a court to undertake "'hands-on' management" of a government agency, *NAACP v. Bureau of the Census*, 945 F.3d 183, 191 (4th Cir. 2019). Defendants reason that happened here because the Census's "data-processing operations, including disclosure avoid-ance, '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." Resp., Doc. 41 at 54 (quoting *NAACP*, 945 F.3d at 191).

But unlike the challengers in *NAACP*, Plaintiffs do not seek "wholesale improvement" of the entire census operation. They request only what they are entitled by law to receive: tabulations of population that accurately reveal the intrastate population counts from the decennial census. *See* 13 U.S.C. § 141(c). To be sure, such an order will have downstream effects. Like any order, it will require Defendants to act, or stop acting, to fulfill the law's mandate. Contra Defendants' intimation, though, there is no too-big-to-fail exception to judicial review. Nor would such an order "inevitably" lead to "hands-on" management of the census. *How* Defendants decide to comply with the order will be up to them. All of that is in stark contrast to what happened in *NAACP*, where the

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 53 of 65

plaintiffs sought to commandeer many aspects of the Bureau's program. Asserting vague harms of "insufficiency"—"insufficient network of area census offices," "insufficient partnership program staffing," "insufficient testing of 'new protocols,'" "insufficient enumerators," 945 F.3d at 190-91—the *NAACP* plaintiffs sought "an injunction that requires Defendants to propose and implement, *subject to this Court's approval and monitoring*, a plan to ensure that hard-to-count populations will be actually enumerated in the decennial census," *NAACP v. Bureau of the Census*, 399 F. Supp. 3d 406, 422 (D. Md.), *aff'd in part, rev'd in part and remanded*, 945 F.3d 183 (4th Cir. 2019) (emphasis added). That is nothing like this case.

Second, Defendants dispute the timing and finality of the action Plaintiffs challenge. On the one hand, they say, the Bureau's decision to use differential privacy is not final because "[c]ritical details of how the Census Bureau will implement differential privacy have not yet been finalized," Resp., doc. 41 at 57, and the Bureau's action will not be final until "the Secretary delivers the final data to the States," *id.* at 56. On the other hand, they argue, Plaintiffs' claims are barred by laches because "Plaintiffs have delayed considerably in asserting" them. *Id.* at 64.

For their too-early argument, Defendants first turn to *Franklin v. Massachusetts*, 505 U.S. 788 (1992), a case involving a challenge to the Bureau's method of allocating federal overseas employees in the apportionment numbers that cost Massachusetts a congressional seat. The Court held that, because "the action that creates an entitlement to a particular number of Representatives and has a direct effect on the reapportionment is the President's statement to Congress, not the Secretary's report to the President," Massachusetts could not challenge the *Bureau's* decision under the APA. *Id.* at 797; *see id.* (noting that agency action is final under the APA if (1) "the agency has completed its decisionmaking," and (2) the "the result of that process is one that will directly affect the parties"). Likening this case to *Franklin*, Defendants thus contend that "[f]inal action

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 54 of 65

will occur only when the Secretary delivers the final data to the States, which has not yet occurred." Resp., Doc. 41 at 56. But this case is not *Franklin*. Unlike with the apportionment counts under § 141(b), there is no higher official that needs to weigh in before the Bureau irreversibly implements differential privacy to the redistricting tabulations under § 141(c). And unlike the President, who can tell the Bureau "to reform the census, even after the data are submitted to him," *id.* at 798, the States cannot tell the Bureau to reform the redistricting data once it is released. *Cf. City of New York v. U.S. Dep't of Com.*, 822 F. Supp. 906, 918 (E.D.N.Y. 1993) (noting that *Franklin* "did not involve a situation where … plaintiffs challenge the counts as they are used for intra-state redistricting and for federal fund allocation (emphasis omitted)), *vacated on other grounds*, 34 F.3d 1114 (2d Cir. 1994), *rev'd sub nom. Wisconsin v. City of New York*, 517 U.S. 1 (1996).

More fundamentally, *Franklin* does not stand for the proposition Defendants seem to glean from it—that, because the *President's* handing over to Congress the apportionment data constituted final agency action (as opposed to the Secretary's delivery of the data to the President), anything short of releasing the redistricting data to the States cannot constitute final agency action. *See* Resp., Doc. 41 at 56. *City of Detroit v. Franklin*—on which Defendants also rely for the same contention—proves the point. *See* 4 F.3d 1367 (6th Cir. 1993). Though noting that the Secretary's reporting redistricting data to the States qualified as final agency action, *id.* at 1377 n.6, the Court actually reviewed a later action: the Secretary's refusal to adjust and re-report the redistricting data, *see id.* at 1377-78. By Defendants' logic, the fact of a later decision would necessarily make the release itself not final and unchallengeable. That, of course, was not so. Neither is it so that any other census-related action short of sharing census data cannot constitute final agency action. It can. *See Dep't of Com. v. New York*, 139 S. Ct. 2551, 2578 (2019) (reviewing under the APA the Bureau's decision to include citizenship question on the census form).

Next, Defendants invoke the changing nature of the differential privacy system and argue that the Bureau's decision cannot be final until it "sets the final privacy-loss budget and releases the final numbers." Resp., Doc. 41 at 58. Until that happens, Defendants say, Plaintiffs cannot show "that there will be *any* legal consequences from differential privacy." *Id.*; *see also id.* at 57 (noting that "[c]ritical details of how the Census Bureau will implement differential privacy have not yet been finalized"). To be sure, it is true that Census officials are still working on differential privacy—that's part of the problem given the March 31 deadline. But it is not true that Plaintiffs will escape harm based on any of the decisions left to make. That is because the Bureau has already "finalized"—its word—"the list of 'invariants' for the first set of 2020 Census data products," which includes "the P.L. 94-171 file (redistricting data)":

# 11/25/2020: Invariants Set for 2020 Census Data Products

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 noise infusion.

The first set of data products includes the P.L. 94-171 file (redistricting data), the Demographic Profiles, the Demographic and Housing Characteristics File, and the Congressional District Summary File. These products will be protected using the main TopDown Algorithm (TDA) central to the Disclosure Avoidance System (DAS). We will employ other formal privacy methods to handle the remaining more detailed data tables separately to preserve the greatest degree of accuracy, as they pose difficult and unique privacy challenges.

Per the decision, the following statistics for this first set of products will be invariant at these levels of geography and higher:

- Total population (at the state and state-equivalents level [1])
- Total housing units (at the census block level)
- Number of group quarters facilities by type (at the census block level)

Ex. 8, U.S. Census Bureau, 2020 Disclosure Avoidance System Updates: Invariants Set for 2020 Census Data Products (Nov. 25, 2020), https://perma.cc/G5WE-S58C.

Thus, as Plaintiffs explained in their Complaint, all other tabulations—"such as how many

people live in a census block"-will be skewed intentionally. Doc. 1 at 20; see also id. at 21 ("The

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 56 of 65

Bureau's adoption of differential privacy will cause the Bureau, for the first time ever, to purposefully report incorrect population counts to the States to use for redistricting."); *id.* at 29-30 ("Congress did not give the Bureau authority to report estimates or values that merely bear some relation to sub-state population counts. Congress required that the actual numbers be reported."). Defendants do not deny that. But it is the actual population counts within the State that Plaintiffs are entitled to under § 141(c). And it is those same counts that Defendants *will* withhold absent action by this Court. There is nothing uncertain about that. The challenged agency action is final and ripe for review under the APA.

As for their too-late argument, Defendants invoke the doctrine of laches to assert that (1) Plaintiffs delayed in bringing their claims for relief, (2) the delay was inexcusable, and (3) the delay caused Defendants undue prejudice. See Resp., Doc. 41 at 63-64. The facts belie such claims. Plaintiffs did not delay, but neither did they rush to court prematurely. It did not become apparent that the Bureau's application of differential privacy would abridge Plaintiffs' rights under § 141(c) until November 24, 2020, when the Data Stewardship Executive Policy Committee set the invariants. Before that day, Plaintiffs "were entitled to presume that the public officials responsible for" producing redistricting data "would act in accordance with the law." Black Warrior Riverkeeper, Inc. v. U.S. Army Corps of Engineers, 781 F.3d 1271, 1285 (11th Cir. 2015). And after that day, Plaintiffs had a "reasonable need to fully investigate [their] claims," id., and assess the gravity of the harm Defendants would inflict, which Plaintiffs promptly did, see Thomas B. Bryan, Census 2020: Differential Privacy Analysis - Alabama Case Study, Doc. 3-6. The short amount of time Plaintiffs took "to properly prepare for" this "complex ... litigation ... cannot constitute inexcusable delay." Black Warrior Riverkeeper, 781 F.3d at 1285. A contrary ruling "would create a powerful and perverse incentive for plaintiffs to file premature and even frivolous suits to avoid the

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 57 of 65

invocation of laches." *Id.* Nor were Defendants unduly prejudiced in any event. If they are violating the law as alleged, the costs of changing course hardly qualifies as undue prejudice in equity.

# **B.** Defendants' Delay Releasing the Tabulations of Population Is Unlawful.

# 1. Section 209 Affords Plaintiffs a Cause of Action Because Differential Privacy Contributed to the Delay.

The Bureau's implementation of differential privacy has at least partly contributed to and will continue contributing to the complained-of delay. Accordingly, Pub. L. No. 105-119, § 209 provides Plaintiffs with a cause of action to challenge Defendants' delay.

As explained above, § 209 broadly applies to "[a]ny person aggrieved by the use of any statistical method in violation of the Constitution or any provision of law." *Id.* § 209(b). As threshold matters, Plaintiffs constitute "person[s] aggrieved," *see* doc. 2 at 6-8, and differential privacy constitutes a "statistical method," *see* docs. 25, 27. And despite Defendants' protestations otherwise, *see* Resp., doc. 41 at 65-66, Plaintiffs do not challenge the February 12 press release as such (as though Plaintiffs' problem is with the press release qua press release), but the decision reflected in the press release: the Bureau's decision to delay delivering the redistricting data. There is ample reason to think that delay is caused, at least in part, by differential privacy.

Defendants contend that differential privacy—though months behind schedule—has nothing to do with overdue redistricting data. Their first argument, however, rests on a sleight of hand. Defendants deny "that applying any other appropriate disclosure-avoidance methods *that complied with Title XIII* would allow them to release Redistricting Data any sooner than using Differential Privacy." Ex. 12, Requests for Admission at 1-2 (emphasis added). Defendants' position seems to be that differential privacy—or something close to it—is the only disclosure-avoidance method that complies with Title XIII's confidentiality requirements. By that logic, the Bureau could spend

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 58 of 65

the next few years trying and failing to implement differential privacy and still deny that differential privacy "has, at least in part, delayed the release" of data. *Id.* at 1. In any event, as explained above, *see* II.A.1, the purported re-identification attack that is the foundation for differential privacy was a failure. Numerous other "appropriate disclosure-avoidance methods that compl[y] with Title XIII" are out there, if the Bureau will only use them. There is every reason to think the Bureau still has those methods at its disposal. Ex. 6, Ruggles Expert Report at 13-16. JASON even recommended that the Bureau "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." *See* JASON, *supra*, at 8, 98. If the Bureau followed that advice, it can necessarily apply swapping more quickly than it can apply swapping and then differential privacy. And even if it rejected JASON's advice, "there should be few barriers to applying the 2010 software to the 2020 data." Ex. 6, Ruggles Expert Report at 13.

Second, issues with differential privacy have plagued the Bureau throughout the 2020 census, suggesting that they—at least in part—contributed to the extensive delay. The Bureau "announced its decision to use differential privacy for the 2020 census in September 2017," Resp., doc. 41 at 64, yet differential privacy is *still* a work in progress. Indeed, just this past September, the Census Scientific Advisory Committee noted its reservations with the program's progress:

The Bureau's implementation of differential privacy has followed an ambitious timeline under any circumstances, even in the absence of a global pandemic or other challenges. The Bureau is operating under enormous time pressure to make the incredibly consequential and irreversible decision on the privacy-loss budget and its allocation. But many implications of this decision for privacy, accuracy, and fitness-for-use are currently unknown. The process by which the Bureau will determine the privacy-loss budget allocation is unclear. Whatever the choice of privacy-loss budget allocation, the Bureau will need to estimate the re-identification risk to ensure sufficient privacy, will need to give users methods for assessing fitness-foruse, and will need to have a backup plan (e.g., allocate some privacy budget) for the future, in case differentially-private data are not fit for some important use cases. The recommended use case catalog development and rigorous analysis for

priority use cases are important for informing how to allocate the privacy-loss budget across uses.

Ex. 13, U.S. Census Bureau, Census Scientific Advisory Committee Fall 2020 Meeting Recommendations 13-14 (Sept. 17-18, 2020), https://perma.cc/RR3S-XEAU. As a result, the Committee recommended delaying "additional releases after the December apportionment release to allow time" to resolve the significant problems with the system. *Id.* at 15.

That was *September*—six months before the redistricting data were to be released. Then, in October, when confronted with concerns that "more folks are asking for the [differential privacy disclosure avoidance system to take a backseat," Abowd replied: "I don't know how to speed up the work. I agree … that we should hold everything after the redistricting data until there is consensus that the underlying data are fit-for-use. *But there's no way to slow the redistricting [data] long enough to get that*." Ex. 14, Email from John Abowd to Danah Boyd (Oct. 7, 2020) (emphasis added). As fortune would have it, four months later the Bureau announced that it would indeed "slow the redistricting" data. If that decision was unrelated, it sure was serendipitous.

Lastly, Defendants attempt to sever the link between differential privacy and delay by asserting that "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." Resp., Doc. 41 at 66. Of course, there is no practical difference between "approximately three weeks" and "nearly four weeks." More significantly, the Bureau initially allotted itself until no later than March 31, 2021, to have differential privacy applied to the redistricting data, but as of today, "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." Bryan Rep., Doc. 3-6 at 10. Reverting to proven disclosure avoidance methods is far more likely to result in

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 60 of 65

expeditious production of data than sticking with an unnecessary and experimental approach that the Bureau can't seem to get right.

# 2. Defendants' Delay Violates the Census Act.

In its February 12 Decision, the Bureau declared it would "deliver the Public Law 94-171 redistricting data to all states by Sept. 30, 2021." Doc. 3-7. This data was due March 31, 2021. 13 U.S.C. § 141(c). The Bureau thus missed the deadline, and, in so doing, violated the Census Act. Defendants do not disagree, and instead argue that Plaintiffs lack a cause of action to sue under the Act. But as explained above and in Plaintiffs' additional briefing on the matter, *see* docs. 2 & 25, § 209 provides Plaintiffs a cause of action. And, as discussed next, Plaintiffs also have a cause of action under the APA. But even if Plaintiffs could not bring suit under § 209 or the APA, this Court's inherent equitable power allows it to enjoin the Bureau from "violating, or planning to violate, federal law." *Armstrong v. Exceptional Child Ctr., Inc.*, 575 U.S. 320, 326-27 (2015).

# *3. Defendants' Delay Violates the Administrative Procedure Act.*

*The February 12 Decision constitutes final agency action*. The Bureau's decision to delay releasing the redistricting data beyond the statutory deadline marked "the consummation of the [Bureau's] decisionmaking process," and was not "of a merely tentative or interlocutory nature." *U.S. Army Corps of Eng'rs v. Hawkes Co.*, 136 S. Ct. 1807, 1813 (2016). In fact, the statutory deadline has already passed, so it is impossible for the Bureau to claim that it is still considering whether to meet it. The "legal consequences" of that decision have begun flowing, too. *See id.* The decision guaranteed that the Secretary would violate Plaintiffs' rights under § 141(c) to receive timely redistricting data, and Plaintiffs have already been harmed by that decision. Both prongs of "final agency action" are thus satisfied.

Defendants present several arguments in response, most of them (like Defendants' misreading of *Franklin*) discussed above. Their main protest here is that "a press release [*i.e.*, the

# Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 61 of 65

February 12 Decision] explaining that the Census expects to deliver redistricting data by a certain date did not consummate anything" because it was merely "snapshot" "updat[ing] Census's estimated timeline, and of course, estimates can still change." Resp., Doc. 41 at 69. But it matters not that the Bureau's decision was made public through a press release, as opposed to a more official looking document. Otherwise an agency could evade APA review for any action as long as it communicated the action through public media. Nor does it matter that the Bureau's Decision is theoretically subject to change. For one, the deadline has passed; federal law has already been violated. For another, the possibility of further change is not the test—indeed, if the APA's finality inquiry turned on whether an agency action *could* someday change, virtually no agency action would ever be final. Instead, "[t]he finality requirement is concerned with whether the initial decisionmaker has arrived at a definitive position on the issue that inflicts an actual, concrete injury." *Darby v. Cisneros*, 509 U.S. 137, 144 (1993). The "issue" here was whether to comply with March 31 deadline; the Bureau "arrived at a definitive position" not to; and the decision "inflict[ed] ... actual, concrete, injur[ies]." *Id.* 

The February 12 Decision was arbitrary and capricious and otherwise not in accordance with law. The February 12 Decision should be set aside because it was arbitrary and capricious. 5 U.S.C. § 706(2)(A). As Plaintiffs initially explained, the Bureau knows that States like Alabama rely on accurate, timely census data for redistricting. Mot, Doc. 3 at 58. Yet the Bureau discounted the States' interests, and did not even attempt to mitigate the harm of delay by producing data on a rolling, or "flow," basis (as it has in the past) or prioritizing States whose laws require timely data. The February 12 Decision thus evinced disregard for the significant reliance interests States have in the timely production of redistricting data, and further revealed a poorly conceived response to the Decision's predictable repercussions. *See Dep't of Homeland Sec.*, 140 S. Ct. at 1915.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 62 of 65

Defendants' response amounts to an admission. Offering one sentence to prove the Decision was not arbitrary or capricious, the Bureau *concedes* it could release data to Alabama *weeks* earlier than it currently plans to, but that doing so would delay other States' receipt of data. Doc. 41 at 71. But that is Plaintiffs' point—the Bureau could have provided the data first to those States who would otherwise suffer injury, and it should have planned to do so. The Bureau's own statements show that the February 12 Decision was arbitrary and capricious, and its minimal arguments to the contrary fail.

# **III.** The Equities Warrant Injunctive Relief.

The equities favor swift relief. Absent an injunction, Plaintiffs will be irreparably harmed. Their statutory and constitutional rights will be violated. The State will be forced to redistrict with tabulations of population that do not accurately reflect the population counts within the State (or, yes, as Defendants repeatedly suggest, the State could spend billions of dollars to build its own census, though that would only exacerbate the delay problem). And the harm will be irreparable because the State needs to begin redistricting promptly and so will be forced to use Bureau's second-rate data upon its release. *See* Mot., Doc. 3 at 61-69. Defendants' rejoinders have been addressed at length above. *See* Resp., Doc. 41 at 73-79. The Court should order relief.

# IV. In The Alternative, Plaintiffs Are Entitled To A Writ Of Mandamus.

Alternatively, the Court should issue a writ of mandamus requiring the Secretary to deliver the redistricting data to Alabama as soon as possible. Defendants protest that the word "shall" in § 141(c) is merely precatory, even as they concede that "the word 'shall' usually connotes a requirement." Resp., Doc. 40 at 80 (quoting *Maine Cmty. Health Option v. United States*, 140 S. Ct. 1308, 1320 (2020) (emphasis omitted)). And they list previous acts of Congress excusing census delays as "support" for their "conclusion that census deadlines are directory in nature," *id.* at 81, 82 n.7, even though the listing suggests just the opposite. (Why did Congress act if it thought the Bureau had all the time in the world?) What is notable, though, is what is missing from that listing: any word from the current Congress or the last excusing the Bureau's delay, despite Defendants' request for an extension. The Secretary's duty is mandatory, and this Court should enforce it.

# **CONCLUSION**

In response to an academic's concerns over the differential privacy system, a research contractor at the Census Bureau warned Abowd that responding with "'You have to trust us' ain't gonna work."<sup>26</sup> Yet that is precisely what Abowd and the Defendants offer here. They promise that the final version of the tabulations—whenever they are released—will be fit for use, and urge the Court to trust them and deny Plaintiffs' requested relief. *See* Abowd Decl., Doc. 41-1 at 47; Resp., Doc. 41 at 20. And, they threaten, the Court *must* trust them because the Bureau has gone all-in on differential privacy on a tightened timeline with no Plan B.

As demonstrated above, these arguments are wrong factually and legally. Plaintiffs do not ask the Court to micro-manage the last phases of Census 2020. But the law entitles Plaintiffs to relief, and it is this Court's duty to provide it. And Defendants have a number of ways to deliver accurate redistricting data well before September. Accordingly, this Court should (1) enjoin Defendants from reporting inaccurate population counts to Plaintiffs, and (2) enjoin Defendants from delaying the release of the P.L. 94-171 redistricting data any longer than necessary, and by no event beyond July 31, 2021.

<sup>&</sup>lt;sup>26</sup> See Ex. 14, Email from Danah Boyd to John Abowd (Oct. 5, 2020).

# Dated: April 20, 2021

STEVE MARSHALL Attorney General of Alabama

<u>/s/ Edmund G. LaCour Jr.</u> 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

Respectfully submitted,

/s/ Jason B. Torchinsky

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

Case 3:21-cv-00211-RAH-ECM-KCN Document 94 Filed 04/20/21 Page 65 of 65

# **CERTIFICATE OF SERVICE**

I hereby certify that on April 20, 2021, I filed with the Court and served on all counsel through the CM/ECF system the foregoing document.

<u>/s Edmund G. LaCour Jr.</u> Counsel for the State of Alabama Case 3:21-cv-00211-RAH-ECM-KCN Document 94-1 Filed 04/20/21 Page 1 of 6

# Exhibit 1

 To:
 John Maron Abowd (CENSUS/ADRM FED)[john maron.abowd@census.gov]; Victoria Velkoff (CENSUS/ADDP

 FED)[Victoria.A.Vefkoff@census.gov]; Cynthia Davis Hollingsworth (CENSUS/EWD FED)[cynthia.davis.hollingsworth@census.gov]

 From:
 James Whitehorne (CENSUS/ADDC FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP

 (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=010B27ED5B5944CDA73EB75917D5A0DD-WHITEHORNE,]

 Sent:
 Wed 9/30/2020 2:54:22 PM (UTC)

 Subject:
 Re: GQ population and the DAS

James Whitehorne (CENSUS/ADDC FED) has shared a OneDrive for Business file with you. To view it, click the link below.

VTD\_Key\_2010Census.zip

# Good morning John -

I apologize for the delay in responding. I am struggling with how to respond because I am worried about our eventual release of the P.L. 94-171 data and its fitness for use as judged by external parties. The role of my office, in addition to managing the Redistricting Data Program, is to represent the interests of the states in regards to redistricting when I am engaged in internal discussions at the Census Bureau. The feedback that I receive from the redistricting community indicates a need for accuracy for the limited set of characteristics contained in the PL dataset, primarily total pop, major race group, ethnicity, voting age, and group quarters counts. I would not be representing the states and the redistricting community well if I did not pass along and stress these concerns. The example I provided was not intended as a "cherry picked anecdote". As a federal agency that provides the data that underpins every American's rights and how those American's are served by their government, I was trying to illustrate the extremes to which we need to cater. We are not a think tank, a university, or a for profit business chasing after deep pocketed clients, we are a service to the country. This is much like the Postal Service who must provide service to every address, regardless of its profitability. We must ensure we are providing data that can be used for the needs of all communities. I will still promote and defend what we end up producing. However, I am trying to underscore the point that we have all heard and that you and I have heard directly from our nation's voting rights enforcement agents in the Voting Section at DoJ, that accurate data is critical especially for small areas. The data must reflect what is seen in the real world because it is used to change how the real world interacts with itself and with its government. This does not mean I do not understand our obligation to protect the public's data, it just appears that in our zeal to protect the data we are harming the very same people we are protecting. This will be my last emotional response but I had to state my thoughts on this at least once. I am willing and happy to continue to work with you and your team in good faith to make DP at Census a success, as I have done since the beginning.

In response to your specific requests/recommendations, I worked with GEO to generate a national VTD key that can be plugged into the 2010 Census P.L. Data and the PPMFs to generate data summaries for the VTD geography for the nation. I only have an office of four people (two geographers, a program manager, and myself) from which to draw support so unfortunately I do not have an analyst to assign to this but I will help personally however I can.

The files in the attached zip file contain the same information but are in two formats: pipe delimited text and SQLlite database

The fields are: BLOCKID = 15 digit unique within nation block code STATEFP = State FIPS COUNTYFP = County FIPS TRACTCE = Tract code BLOCKCE = Block code VTDST = VTD Code VTDID = Unique within nation VTD code (please note in 2010 this is a variable length code) NOTES: 1) The VTDs in California should be disregarded. They would skew any results because they are not actual VTDs. The

Director allowed an experiment for CA in 2010 where the state submitted an amalgam of multiple district types as pieces that could be reaggregated. Based on the results, we did not allow this for the 2020 Census.

2) It would be wise to keep track of the number of blocks in a VTD when doing the analysis. In some states there can be "island" VTDs. An example of how these occur are situations where a piece of land is surrounded but not part of an

incorporated place If the state has a centiquity requirement for their VIDs, that space has to become a stand alone VTD. These would usually only be a few blocks at most so isolating them from an analysis seems wise. 3) For those counties without VTDs (none submitted to the Bureau) the VTDST code in the key file has been set to ####### and the VTDID code is STATEFP++COUNTYFP++######

Respectfully, James

P.S. Thank you for the good news on the GQ spillover problem in this morning's privacy meeting.

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Monday, September 28, 2020 9:51 AM
To: James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>; Victoria Velkoff (CENSUS/ADDP FED)
<Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/EWD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: GQ population and the DAS

James,

I appreciate the communications problem that transparency about the DA process for the 2020 Census has created for you. I do think we need to put this in proper perspective, and that requires comprehensive statistical summaries, not cherry-picked anecdotes.

We do have the data to produce a summary of the size of every political entity for which we collected district boundary information. If you assign an analyst, I will assign a SME, and together we can get these data produced. Once we have the empirical distribution of voting district sizes, we can use The Wright and Irimata framework to construct fitness for use summaries by expected size of the voting district. This will vary by size of the voting district and PLB assigned to hispanic + cenrace. That has to be the basis for DSEP's decision-making here.

Thanks,

John M. Abowd, PhD, Associate Director and Chief Scientist Research and Methodology U.S. Census Bureau O: <u>301-763-5880</u> M: simulring on cell <u>census.gov</u> | <u>@uscensusbureau</u> Shape your future. START HERE > <u>2020census.gov</u>

From: James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>
Sent: Sunday, September 27, 2020 12:06 PM
To: John Maron Abowd (CENSUS/ADRM FED) <John.maron.abowd@census.gov>; Victoria Velkoff (CENSUS/ADDP FED)
<Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/EWD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: GQ population and the DAS

John -

Thank you for the quick and detailed reply. I want to make sure there is no confusion coming from my forwarded email. I

am not requesting that GQ populations become invariant. My goal with this massage way to make you and Tori aware that the GQ population and accuracy around the GQ population is also an important factor for redistricting. This is especially true this decade due to the focus on GQ populations in the process by several state legislatures. The "spillover" becomes important due to the divergence of the prison population characteristics and the non-prison populations that surround them. This is compounded in many cases where the non-prison populations are often sparse themselves (rural areas). My hope is that this can be considered and provided the appropriate level of PLB to not only provide fit for use data but also that on its face appears fit for use by casual data users.

You substantive point is very difficult to address. It would require cataloging the size of all electoral districts in the country, which are all user defined areas the majority of which the Bureau doe not collect. This drives the need for block level data. To tray and get a sense of this problem. if we look at the extremes on the county level we have (pop estimates from 2017):

Loving County, TX POP = 134 people; County Judge Commissioner Districts = 4; Average district size = 33.5 people King County, TX POP = 296; County Judge Commissioner Districts = 4; Average district size = 74 people Kenedy County, TX POP = 417; County Judge Commissioner Districts = 4; Average district size = 104.25 people Arthur County, NE POP = 457; County Commissioner Districts = 3; Average district size = 152.33 people Blaine County, NE POP = 482; County Commissioner Districts = 3; Average district size = 160.66 people McPherson County, NE POP = 499; County Commissioner Districts = 3; Average district size = 166.33 people Petroluem County, MT POP = 523; County Commissioner Districts = 3; Average district size = 174.33 people Yakutat Borough, AK POP = 605; Assembly members = 5 (plus mayor and mayor pro tem for 7 total); Average district size = 121 people (could not confirm that Boroughs elect from districts rather than at large). Loup County, NE POP = 609; County Commissioner Districts = 3; Average district size = 203 people

These are just the smallest counties in the country. We also have to consider the smallest places that conduct district based elections. If we are to make the claim that the data is fit for use, and we have defined that use as redistricting under the applicable federal statutes covering voting rights, then we must find a way to provide plausibly accurate data for both the total population and for the underlying relevant characteristics. I am not suggesting breaking the DAS, I am suggesting that the accuracy versus privacy trade-off is farther to the accuracy end-point than you are likely comfortable in recommending and that the tuning of the characteristics should be a continuing focus for the DAS.

Thank you James

\*\*\*\*\*

James Whitehorne, Chief Redistricting & Voting Rights Data Office/ADDC/HQ U.S. Census Bureau O: 301-763-4039 | M: 202-263-9144 <u>census.gov</u> | <u>census.gov/rdo | @uscensusbureau</u> Shape your future. START HERE > <u>2020census.gov</u>

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Friday, September 25, 2020 12:36 PM
To: James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>; Victoria Velkoff (CENSUS/ADDP FED)
<Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/EWD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: GQ population and the DAS

I clarified the distinction between unit count and population count invariants multiple times until Cynthia finally said she

understood it. I had no idea that you did not did not did not conscible to make the GO populations invariant. That is non-starter regardless of DSEP's decisions in October. It will break the DAS. As you know, the fact that the prisons know their populations does not excuse the Census Bureau from protecting the number reported to GQE under Title 13, Section 9. That's not the same as the public knowledge that that living quarter is prison, not a dormitory.

That said, I immediately opened two issues with the DAS science team following the FSCPE meeting: single-sex GQs, and "characteristic spillover." Correct me if I am wrong, but single-sex does not affect redistricting, so I have given that one a lower priority. Characteristic spillover is a feature of top-down, not a bug, but the team understands that characteristics in a GQ like prison are not similar to the characteristics in surrounding blocks. We have a fix under investigation (grouping GQs into their own special block-groups).

I note for the record that household racial composition is not a characteristic in PL94-171, so I do not understand that argument.

Now for the most substantive point. We know how to get the populations very, very accurate without imposing invariants (+/- < 1 person at the block level MAE). The DAS team will engineer the PLB allocations as instructed; so, this can be implemented. The problem with doing this is that it destroys the confidentiality protection on residential location, which was the biggest factor making re-identification in the 2010 data so high. I am very uncomfortable recommending a PLB allocation to population that removes all of the location protection. What we need is proper guidance on the correct use case: minimum population in a voting district. My view is that the data do not support voting districts the size of 10 persons, or even 100, and never did. But I can be persuaded by data.

Once the PLB on population is pumped up, the race/ethnicity data will be more noisy than if they had gotten more of the PLB, but Tommy and Kyle's analysis suggests that for the DoJ use cases, once the global PLB is above 3 on the PL94-171 data, this variability does not make much difference. So, there's my data, and we will certainly repeat that whole analysis to confirm that the conclusions still hold once the 2020 DAS for redistricting data is locked algorithmically.

Hope this addresses your concerns (but guessing that there will be more discussion),

John M. Abowd, PhD, Associate Director and Chief Scientist Research and Methodology U.S. Census Bureau O: <u>301-763-5880</u> M: simulring on cell <u>census.gov</u> | @uscensusbureau Shape your future. START HERE > <u>2020census.gov</u>

From: James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>
Sent: Friday, September 25, 2020 12:14 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Victoria Velkoff (CENSUS/ADDP FED)
<Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/EWD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: GQ population and the DAS

# Hello John, Cynthia, and Tori,

I am starting to get repeated questions regarding the treatment of GQ population by the TDA. Some of this is coming from the issues pointed out at the FSCPE meeting and some are coming from review of the latest PPMF. NCSL followed up with the following regarding GQ populations:

What's new, as I understand it, is that DAS means that the population characteristics are more mixed than ever—that the households look far more racially diverse than they otherwise would, at the least.

I've just heard that the new PPMF shows that not only are the characteristics farther from the actual enumeration, but that the actual population count is also treated with DAS and therefore not accurate. (The accurate population count in the prisons can be known by asking the prison officials—in fact, that's how the bureau gets it in the first place.)

I believe vou've severe saving all along that the doff GO and the type of GO and the type of GO will be as enumerated of didn't quite grasp when you've said that the implication: the population count for GQs will not be as enumerated.

First, do I have the issue understood correctly enough? Second, what's a good approach for us to ensure that the DAS team knows that these #s are going to matter a lot to redistricters in 9 states at the very least, and that as they continue fixing the post-processing and set the epsilon, this is a significant concern on the redistricting side? My purpose is two-fold. One is to make sure you are aware that this is a concern for the redistricting use case. Two, to suggest that a comprehensive write up on how GQs are being treated in the TDA and what people should be aware of who need to use the GQ data would be useful. Would this be a good topic for an upcoming newsletter after the invariants are set? Maybe as part of a newsletter on the invariants?

Finally, will there be a public response to the FSCPE questions regarding this topic? Thanks

James

\*\*\*\*\*

James Whitehorne, Chief Redistricting & Voting Rights Data Office/ADDC/HQ U.S. Census Bureau O: 301-763-4039 | M: 202-263-9144 census.gov | census.gov/rdo | @uscensusbureau Shape your future. START HERE > 2020census.gov Case 3:21-cv-00211-RAH-ECM-KCN Document 94-2 Filed 04/20/21 Page 1 of 31

# Exhibit 2


### STATISTICAL EXPERTISE & GENERAL RESEARCH TOPICS

**CENTER FOR STATISTICAL RESEARCH & METHODOLOGY** 

Research & Methodology Directorate U.S. Bureau of the Census (FY 2021 – FY 2025)

TOTALLY OR PARTIALLY FUNDED BY
• WORKING CAPITAL FUND / GENERAL RESEARCH PROJECT

SEPTEMBER 2020



To help the Census Bureau continuously improve its processes and data products, general research activity is undertaken in seven broad areas of statistical expertise and general research topics. The activities are supported primarily by the General Research Project of the Working Capital Fund and results from these activities benefit all (decennial, demographic, and economic) programs as well as advance general statistical methodology and practice. With this update, we have modified the names for some of the areas of expertise to better reflect change and expertise/interest of current staff as well as changing Census Bureau needs. We have also added a new effort focusing on cross-cutting statistical general research priorities where we will form small study/working groups (not necessarily limited to folks in our center) where there seems to be overlap and a real need to look forward.

Expertise for Collaboration and Research <sup>1</sup>		Page
1.	Missing Data & Observational Data Modeling	1
2.	Record Linkage & Machine Learning	4
3.	Small Area Estimation	7
4.	Sampling Estimation & Survey Inference	10
5.	Time Series & Seasonal Adjustment	15
6.	Experimentation, Prediction, & Modeling	19
7.	Simulation, Data Science, & Visualization	22
8.	Cross-cutting Statistical General Research Priorities	25

<sup>&</sup>lt;sup>1</sup>The Center for Statistical Research & Methodology reviews all research activities and results to ensure that *Census Bureau Statistical Quality Standards* are met and that

<sup>•</sup> each effort meets a business need of the Census Bureau (motivation, research problem(s), potential applications), which includes how it aligns with the Census Bureau's strategic plan and the R&M Directorate portfolio management;

<sup>•</sup> each effort is deeply based in the scientific method and the foundations of statistical science; and

<sup>•</sup> each effort does not put at risk the Census Bureau's mission and reputation.

#### Missing Data & Observational Data Modeling

**Motivation:** Missing data problems are endemic in the conduct of statistical experiments and data collection operations. The investigators almost never observe all the outcomes they had set to record. When dealing with sample surveys or censuses, this means that individuals or entities in the survey omit to respond, or give only part of the information they are being asked to provide. Even if a response is obtained, the information provided may be logically inconsistent, which is tantamount to missing. Agencies need to compensate for these types of missing data to compute official statistics. As data collection becomes more expensive and response rates decrease, observational data sources such as administrative records and commercial data provide a potential effective way forward. Statistical modeling techniques are useful for identifying observational units and/or planned questions that have quality alternative source data. In such units, sample survey or census responses can be supplemented or replaced with information obtained from quality observational data rather than traditional data collection. All these missing data problems and associated techniques involve statistical modeling along with subject matter experience.

#### **Research Problem:**

- Correct quantification of the reliability of estimates with imputed values, as variances can be substantially greater than that computed nominally. Methods for adjusting the variance to reflect the additional uncertainty created by the missing data.
- Simultaneous imputation of multiple survey variables to maintain joint properties, related to methods of evaluation of modelbased imputation methods.
- Integrating editing and imputation of sample survey and census responses via Bayesian multiple imputation and synthetic data methods.
- Nonresponse adjustment and imputation using administrative records, based on propensity and/or multiple imputation models.
- Development of joint modeling and imputation of categorical variables using log-linear models for (sometimes sparse) contingency tables.
- Statistical modeling (e.g. latent class models) for combining sample survey, census and/or alternative source data.
- Statistical techniques (e.g. classification methods, multiple imputation models) for using alternative data sources to augment or replace actual data collection.

#### **Current Subprojects:**

- Data Editing and Imputation for Nonresponse (Thibaudeau, Morris, Shao)
- Imputation and Modeling using Observational/Alternative Data Sources (Morris, Thibaudeau)

#### **Potential Applications:**

- Modeling approaches for integrating Economic Census editing and imputation processing, and developing multiple synthetic industry-level Economic Census micro-data.
- Modeling approaches for using administrative records in lieu of or to supplement Decennial Census field visits due to imminent and future design decisions.
- Adapt survey questions in the American Community Survey based on modeling of administrative record quality.
- Produce multiply imputed, synthetic and/or composite estimates of more geographical granular and timely economic activity based on third party data.

#### Accomplishments (October 2018-September 2020):

- Researched, adapted, and implemented nonparametric Bayesian hierarchical models developed by Kim et al. (2017) for integrating Economic Census editing and imputation processing with developing multiple synthetic industry-level Economic Census micro-data that can be publicly shared in place of suppressed estimates.
- Collaborated in adapting an R package based on Kim et al. (2017) to be specifically tailored to edit and multiply impute Economic Census data, and documented specifications in a user's guide.
- Developed multiple synthetic generators to produce industry-level Economic Census micro-data.
- Collaborated to develop Bayesian multiple imputation models for using third party data to produce geographically granular and timely retail sales experimental estimates.
- Applied and completed evaluation of optimization methods for raking balance complexes in the Quarterly Financial Report (QFR) when items can take negative values.
- Showed how to use log-linear models coupled with complementary logistic regression to improve the efficiency (reducing the sampling error) of estimates of gross flows and month-to-month proportions classified by demographic variables. Illustrated methodology on labor force measurements and gross flows estimated from the Current Population Survey.

#### Short-Term Activities (FY 2021):

- Continue researching modeling approaches for using administrative records in lieu of Decennial Census field visits due to imminent design decisions.
- Continue to investigate the feasibility of using third party ("big") data from various available sources to supplement and/or enhance retail sales estimates.
- Continue research, implementation, and resolution of editing and data issues when applying non-parametric Bayesian editing methods to edit and multiply impute Economic Census data.
- Continue research on integration of Bayesian editing and multiple imputation processing with disclosure avoidance and data synthesis processing.
- Extend the analysis and estimation of changes in the labor force status using log-linear models coupled with matching logistic regression methods to the Current Population Survey.
- Research novel categorical distributions for contingency table modeling and joint imputation of categorical variables particularly for clustered data.
- Continue research on accounting for observed zero cells in loglinear models for sparse contingency tables.

#### Longer-Term Activities (beyond FY 2021):

- Joint modeling of response propensity and administrative source accuracy.
- Research practical ways to apply decision theoretic concepts to the use of administrative records (versus personal contact or proxy response) in the Decennial Census.
- Further development of joint administrative record and imputation modeling based on latent class models.
- Research imputation methods for a Decennial Census design that incorporates adaptive design and administrative records to reduce contacts and consequently increases proxy response and nonresponse.
- Joint models of attrition (or response rate) and clustered categorical outcomes using shared random effects with innovative GLMM computational techniques.
- Extend small area estimation modeling for longitudinal data (survey and/or third party) in presence of attrition and/or other type of missing data using log-linear models in tandem with logistic regression.

#### Selected Publications:

- Dumbacher, B., Morris, D.S., and Hogue, C. (2019). "Using Electronic Transaction Data to Add Geographic Granularity to Official Estimates of Retail Sales," *Journal of Big Data*, 6(80).
- Keller, A., Mule, V.T., Morris, D.S., and Konicki, S. (2018). "A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 Census," *Journal of Official Statistics*, 34(3): 1-27.
- Winkler, W. E. (2018). "Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Edit/Imputation," *Research Report Series (Statistics #2018-05)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Morris, D. S. (2017). "A Modeling Approach for Administrative Record Enumeration in the Decennial Census," *Public Opinion Quarterly: Special Issue on Survey Research, Today and Tomorrow, 81(S1): 357-384.*
- Thibaudeau Y., Slud, E., and Gottschalck, A. O. (2017). "Modeling Log-Linear Conditional Probabilities for Estimation in Surveys," *Annals of Applied Statistics 11(2)*, 680-697.
- Morris, D.S., Keller, A., and Clark, B. (2016). "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census," *Statistical Journal of the International Association for Official Statistics*, *32*(2): 177-188.
- Thibaudeau, Y. and Morris, D.S. (2016). "Bayesian Decision Theory to Optimize the Use of Administrative Records in Census NRFU," *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association.
- Bechtel, L., Morris, D.S., and Thompson, K.J. (2015). "Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study." In FCSM Proceedings. Washington, DC: Federal Committee on Statistical Methodology.
- Garcia, M., Morris, D.S., and Diamond, L.K. (2015). "Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on Economic Census Products." *Proceedings of the Joint Statistical Meetings*.
- Winkler, W. and Garcia, M. (2009). "Determining a Set of Edits," *Research Report Series (Statistics #2009-05)*, Statistical Research Division, U.S. Census Bureau, Washington, D.C.
- Winkler, W. E. (2008). "General Methods and Algorithms for Imputing Discrete Data under a Variety of Constraints," *Research Report Series (Statistics #2008-08)*, Statistical Research Division, U.S. Census Bureau, Washington D.C.
- Thibaudeau, Y. (2002). "Model Explicit Item Imputation for Demographic Categories," Survey Methodology, 28(2), 135-143.

Contact: Darcy Morris, Yves Thibaudeau, Jun Shao, Eric Slud, Xiaoyun Lu

Funding Sources for FY 2021:0331 – Working Capital Fund / General Research Project<br/>Various Decennial, Demographic, and Economic Projects

#### **Record Linkage & Machine Learning**

**Motivation:** Record linkage continues to grow in importance as a fundamental activity in statistical agencies. The number of available administrative lists and commercial files has grown exponentially and present statistical agencies with opportunities to accumulate information through record-linkage to support the production of official statistics. In addition to cost, new obstacles to traditional data collection have emerged in the form of possibly recurrent pandemics. These circumstances further motivate the accumulation of information by linking public, private and administrative files. Thibaudeau (2020) describes the strides the Census Bureau, a pioneer in record linkage, has made over the years. While this is impressive, more is needed. With its own suite of in-house record-linkage software packages, such as the "SAS (PVS) Matcher", "BigMatch", "d-blink" and "MAMBA", and easy access to open-source packages, such as "fastLink" and "RecordLinkage in R", the Census Bureau now has access to a wide spectrum of methodologies and the potential to rapidly develop and integrate new ones. The Census Bureau must remain abreast of the ever improving state-of-the-art in record linkage and be prepared to champion its own methodologies as some of the best in the world. Our goal is to achieve the synergy of methods and software that will benefit most the Census Bureau and its mission. System portability is also an objective. The Census Bureau should have the freedom to upgrade its IT infrastructure knowing record-linkage applications will remain functional.

#### **Research Problems:**

One challenge is continuing to research and experiment with new methodologies on multiple software platforms while also moving toward integration. Description of such experiments are:

- Markov Chains Monte-Carlo (MCMC), like that powered by d-blink, give full probabilistic characterizations of the recordlinkage process and are becoming indispensable for full comprehension of a record linkage process. At the same time MCMCs can be tweaked to deliver fast snapshots of the linked population. Research in that direction is crucial. Old-School programs like BigMatch have been greatly optimized for fast linking but lack in nuance. They need to be garnished by richer comparison schemes, such as dictionary-assisted fuzzy string comparisons.
- New data structure for record-linkage of multiple large lists need to be explored. d-blink is an example of a more efficient data structure: Node-connected structures minimize the number of comparisons, as opposed to a traditional all pairwise comparisons. Other structures are possible, such as cyclical linked lists (Thibaudeau 1992), and should be researched.
- As new techniques continue to be implemented and experimented on various existing software (R, Python, C) and hardware (Windows, OSX, IRE, CAES) platforms, the dominant paradigms are emerging and work toward integration and unification, while maintaining versatility, is moving in high gear.

#### **Current Subprojects:**

- Adjusting the Statistical Analysis on Integrated Data (Ben-David)
- Entity Resolution and Merging Noisy Databases (Steorts, Brown/CES, Blalock/CODS, Thibaudeau)
- Record-Linkage Support for the Decennial Census (Ben-David, Weinberg, Brown/CES, Thibaudeau)

#### **Potential Applications:**

- Possible massive concurrent record-linkage implementations for Census 2030. The objective is counting all distinguishable persons in linked and unduplicated administrative and commercial person-level lists.
- Unduplication and record-linkage for frame construction in the demographic and economic areas.
- Re-identification through record-linking for proofing confidentiality of data lists.
- Analysis and estimation based on linked lists.
- Linking probabilistic design-based surveys to large non-probability lists and sample for probabilistic calibration.

#### Accomplishments (October 2018-September 2020):

- Deployed the FEBRL (Peter Christen) Python file simulator on multiple platforms.
- Used FEBRL to simulate candidate files –up to 500k records each- for record-linkage with known "true-links". Used the simulated files to assess the rates of precision and recall of "BigMatch" and the "SAS (PVS) matcher".
- Presented poster at Data Linkage Day 2019 entitled: "False Duplicates in the Census: A Novel Approach to identifying False Matches from Record Linkage Software"

- Wrote EM algorithm for estimating the weights of the Fellegi-Sunter record-linkage model for use with "BigMatch" and the R "RecordLinkage" package.
- Installed BigMatch on multiple platforms IRE, Windows, MacOS (simulated data).
- Ran experiments to measure and compare the performance in speed and CPU cycles of a multi-core linkage software (R fastLink) and an optimized single-core record-linkage software (BigMatch) in a multi-core environment.
- Used BigMatch for multiple linkage projects, including the linkage of commercial files, in the construction of a master reference file at the person and housing unit levels for research and experimentation in preparation for Census 2030.

#### Short-Term Activities (FY 2021):

- Provide advice to individuals who plan to update and maintain the programs for record linkage and related data preparation.
- Conduct research on record linkage error-rate estimation, particularly for unsupervised and semi-supervised situations.
- Evaluate "R" vs "Python" packages for record linkage focusing on fuzzy string comparison.
- Assess the possibility of using a surname and given-name reference directory for record-linkage in decennial-census production.
- Continue to research statistical and data-science methods for record record linkage. Explore and compare in-house and "off-the-shelf" packages implementing these methods. Ascertain the competency of record-linkage methods at the Census Bureau.
- Extending record linkage outside the PIK universe.

#### Longer-Term Activities (beyond FY 2021):

- Construct census-based equivalence dictionaries of U.S. given names and surnames for cross-referencing and supervised learning in record-linkage.
- Integrate new methods in our in-house record-linkage engines. Consider the integration of off-the-shelf packages when advantageous.
- Evaluate and compare in-house and off-the-shelf data-science programs and packages (R and/or Python) to construct engines for massive numbers of record-linkage runs for Census 2030.
- Further develop Markov Chain Monte-Carlo applications embedding record-linkage methods in massive parallel processing. Develop methods for extracting record-linkage snapshots from MCMCs.

#### **Selected Publications:**

- Thibaudeau, Y. (In progress). "New Record Linkage Solutions for Demographic Methods at the Census Bureau," Research Report Series (Statistics #2020-??), Center for Statistical Research & Methodology, U.S. Census Bureau, Washington, D.C.
- Slawski, M. and Ben-David, E. (2019). "Linear Regression with Sparsely Permuted Data," *Electronic Journal of Statistics*, Vol 13, No. 1, 1-36.
- Slud, E. and Thibaudeau, Y. (2019). "Multi-outcome Longitudinal Small Area Estimation A Case Study," *Statistical Theory and Related Fields*, DOI: 10.1080/24754269.2019.1669360.
- Steorts, R.J., Tancredi, A., and Liseo, B. (2018). "Generalized Bayesian Record Linkage and Regression with Exact Error Propagation" in Privacy in Statistical Databases (Lecture Notes in Computer Science 11126) (Eds.) Domingo-Ferrer, J. and Montes, F., Springer, 297-313.
- Steorts, R.J. and Shrivastava, A. (2018). "Probabilistic Blocking with an Application to the Syrian Conflict," in Privacy in Statistical Databases (Lecture Notes in Computer Science 11126) (Eds.) Domingo-Ferrer, J. and Montes, F., Springer, 314-327.
- Winkler, W. E. (2018). "Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Editing/Imputation," in (A.Y. Chun and M. D. Larsen, eds.) Administrative Records for Survey Methodology, J. Wiley, New York: NY.
- Thibaudeau, Y., Slud, E., and Gottshalck, A. (2017). "Log-Linear Conditional Probabilities for Estimation in Surveys," *Annals of Applied Statistics*, *11*, 680-697.
- Czaja, W., Hafftka, A., Manning, B., and Weinberg, D. (2015). "Randomized Approximations of Operators and their Spectral Decomposition for Diffusion Based on Embeddings of Heterogeneous Data," 3rd International Workshop on Compressed Sensing Theory and Its Applications to Radar, Sonar and Remote Sensing (CoSeRa).
- Winkler, W. E. (2015). "Probabilistic Linkage," in (H. Goldstein, K. Harron, C. Dibben, eds.) *Methodological Developments in Data Linkage*, J. Wiley: New York.

- Weinberg, D. and Levy, D. (2014). "Modeling Selective Local Interactions with Memory: Motion on a 2D Lattice," *Physica D* 278-279, 13-30.
- Winkler, W. E. (2014a). "Matching and Record Linkage," Wiley Interdisciplinary Reviews: Computational Statistics, http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WICS1317.html, DOI:10.1002/wics.1317, available from author by request for academic purposes.
- Winkler, W. E. (2014b). "Very Fast Methods of Cleanup and Statistical Analysis of National Files," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM.

Winkler, W. E. (2013). "Record Linkage," in Encyclopedia of Environmetrics. J. Wiley.

Winkler, W. E. (2013). "Cleanup and Analysis of Sets of National Files," Federal Committee on Statistical Methodology, Proceedings of the Bi-Annual Research Conference, http://www.copafs.org/UserFiles/file/fcsm/J1\_Winkler\_2013FCSM.pdf., https://fcsm.sites.usa.gov/files/2014/05/J1\_Winkler\_2013FCSM.pdf

- Winkler, W. E. (2011). "Machine Learning and Record Linkage" in Proceedings of the 2011 International Statistical Institute.
- Herzog, T. N., Scheuren, F., and Winkler, W. E. (2010). "Record Linkage," in (Y. H. Said, D. W. Scott, and E. Wegman, eds.) Wiley Interdisciplinary Reviews: Computational Statistics.
- Winkler, W. E. (2010). "General Discrete-data Modeling Methods for Creating Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties," http://www.census.gov/srd/papers/pdf/rrs2010-02.pdf.
- Winkler, W. E., Yancey, W. E., and Porter, E. H. (2010). "Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Alexandria, VA.
- Winkler, W. E. (2009a). "Record Linkage," in (D. Pfeffermann and C. R. Rao, eds.) Sample Surveys: Theory, Methods and Inference, New York: North-Holland, 351-380.
- Winkler, W. E. (2009b). "Should Social Security numbers be replaced by modern, more secure identifiers?", *Proceedings of the National Academy of Sciences*.
- Alvarez, M., Jonas, J., Winkler, W. E., and Wright, R. "Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project," *Electronic Voting Technology*.
- Winkler, W. E. (2008). "Data Quality in Data Warehouses," in (J. Wang, Ed.) *Encyclopedia of Data Warehousing and Data Mining (2<sup>nd</sup> Edition)*.

Herzog, T. N., Scheuren, F., and Winkler, W. E. (2007). Data Quality and Record Linkage Techniques, New York, NY: Springer.

- Yancey, W. E. (2007). "BigMatch: A Program for Extracting Probable Matches from a Large File," Research Report Series (Computing #2007-01), Statistical Research Division, U.S. Census Bureau, Washington, D.C.
- Winkler, W. E. (2006a). "Overview of Record Linkage and Current Research Directions," *Research Report Series (Statistics #2006-02)*, Statistical Research Division, U.S. Census Bureau, Washington, D.C.
- Winker, W. E. (2006b). "Automatically Estimating Record Linkage False-Match Rates without Training Data," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, CD-ROM.
- Yancey, W. E. (2005). "Evaluating String Comparator Performance for Record Linkage," *Research Report Series (Statistics #2005-05)*, Statistical Research Division, U.S. Census Bureau, Washington, D.C.
- Thibaudeau, Y. (2002). "Model Explicit Item Imputation for Demographic Categories," Survey Methodology, 28, 135-143.

Thibaudeau, Y. (1993). "The Discrimination Power in Dependency Structure in Record Linkage," Survey Methodology, 19, 31-38

- Thibaudeau, Y. (1992). "Identifying Discriminatory Models in Record Linkage," Proceedings of the Section on Statistical Computing, American Statistical Association, Alexandria, VA.
- Winkler, W. and Thibaudeau, Y. (1991). "An Application of the Fellegi-Sunter Model of RecordLinkage to the 1990 Decennial Census," *Research Report Series (Statistics) RR91/09*, Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Contact: Yves Thibaudeau, Edward H. Porter, Emanuel Ben-David, Dan Weinberg

 Funding Sources for FY 2021:
 0331 – Working Capital Fund / General Research Project

 Various Decennial, Demographic, and Economic Projects

#### **Small Area Estimation**

**Motivation:** Small area estimation is important in light of a continual demand by data users for finer geographic detail of published statistics and for various subpopulations. Traditional demographic sample surveys designed for national estimates do not provide large enough samples to produce reliable direct estimates for small areas such as counties and even most states. The use of valid statistical models can provide small area estimates with greater precision; however, bias due to an incorrect model or failure to account for informative sampling can result.

#### **Research Problems:**

- Development of models that combine data across multiple surveys or combines survey and observational data (non-probability samples) to improve survey estimates.
- Development of model diagnostic and model comparison tools for small area models.
- Development of small area share models for subareas estimates (e.g. school districts or tracts).
- Development of a design-based simulation system which mimics the American Community Survey to use as a test-bed for area- and unit-level small area models, estimation (both model-based and design-based) methodology and estimation of uncertainty measures.
- Study of measurement error in small area estimation models.
- Development of temporal small area estimation techniques.
- Development of spatial small area estimation techniques.
- Development of more robust estimates of mean squared error of prediction by incorporating Bayesian and bootstrap methods.
- Development of unit-level model framework which appropriately takes into account the complex design of the survey.

#### **Current Subprojects:**

- Using ACS Estimates to Improve Estimates from Smaller Surveys via Bivariate Small Area Estimation Models (Franco, Bell/R&M)
- Bootstrap Mean Squared Error Estimation for Small Area Means under Non-normal Random Effects (Maples, Datta, Irimata, Slud)
- Developing correlated small area share models to create estimates of school district child poverty and population (Maples)
- Developing graphical methods to assess the assumption of constant parameter values across all domains (Maples, Dompreh)
- Developing Bayesian pseudolikelihood models for unit-level data obtained from a complex sample survey (Janicki)
- Assessment of mean squared errors of empirical best linear unbiased predictors for misspecified models (Datta, Slud)

#### **Potential Applications:**

- Borrowing strength from ACS estimates using bivariate modeling has many potential applications, including improving estimates from smaller surveys such as SIPP, NHIS, and CPS, and improving the ACS one year estimates themselves using the previous ACS 5-year estimates.
- Model diagnostic and comparison tools can be applied in any small area application, from SAIPE to SAHIE, to small area models applied to SIPP, AHS, etc.
- The design-based simulation framework for evaluating modes can be used for SAIPE, SAHIE, and other small area programs that use ACS data. The framework can also test the properties of design-based/assisted estimation procedures, such as improvements of sampling variance estimates, propensity score models etc.
- Temporal extensions of small area models will be potentially useful in the VRA Section 203B determinations, and can be applied to ACS data in general, as well as to other surveys that are repeated over time.
- The evaluation of measurement error will help determine if it is appropriate to use ACS-estimates as covariates in models for the Section 203B determinations, and at what level of aggregation.
- Small area share models may be a replacement to the current for the current school district estimates procedures for SAIPE.
- Spatial small area models can improve estimates and provide limited disclosure avoidance for some of the ACS special tabulations.

#### Accomplishments (October 2018-September 2020):

- Developed empirical and theoretical evidence that shows the strong potential of borrowing strength from ACS estimates to improve estimates from smaller U.S. sample surveys using simple bivariate small area estimation models, including applications to NHIS and SIPP, and an application that improves ACS one-year estimates using previous five-year estimates.
- Developed a small area share model to estimate the number of school aged children in poverty for school districts given the official county level poverty estimates.
- Studied alternative models for SAIPE county estimates of school-aged children in poverty using a design-based simulation, and explored the impact of sampling variance estimation in model selection, exploring how design-based estimate and GVF-based estimates impact performance.
- Derived several different mean squared error estimators, both analytical and bootstrapped-based, which will be evaluated in a large simulation study.
- Studied the impact of differential privacy noise infusion on voting district plans and evaluated measures of variability.

#### Short-Term Activities (FY 2021):

- Extend the Small Area Shares model to allow for dependence between sets of shares, e.g. allow the school district to county shares of school age children in poverty and not-in-poverty to have a dependence.
- Finish creating the Artificial Population which mimics the distribution of the U.S. population and implement an ACS-like survey design.
- Improve predictions in ACS special tabulations using a mixture of spatial models.
- Evaluate different mean squared error estimates under the Fay-Herriot model when the error distribution is not always correctly specified.
- Study the impact of measurement error in covariates in small area models for the Voting Rights Act Section 203 determinations.
- Explore times series extensions of the Multinomial Logit model and determine suitability for Voting Rights Act Section 203 determinations.
- Develop multivariate spatial models which use differentially private measurements and auxiliary survey data for the purpose of predicting the number of persons in counties and AIAN areas for detailed race groups'.

#### Longer-Term Activities (beyond FY 2021):

- Develop graphical methods to test assumptions about constant model parameters across all areas.
- Develop models that jointly model survey-weighted proportions and effective sample sizes.
- Explore if a time series model can be applied to improve sampling variance estimates by borrowing strength from estimates from previous years.
- Evaluation of new models (county and school district) to update official SAIPE methodology.
- Deliver a set of 1000 independent survey samples from the Artificial Population with a design similar to the American Community Survey.

#### Selected Publications:

- Erciulescu, A., Franco, C., and Lahiri, P. (2020). "Use of Administrative Records in Small Area Estimation." To appear in Chun, A. Y. and Larsen, M. (Eds.) *Administrative Records for Survey Methodology*. New York: Wiley.
- Parker, P. A., Holan, S. H., and Janicki, R. (2020). "Bayesian Unit-Level Modeling of Count Data under Informative Sampling Designs," Stat, 9.
- Maples, J. (2019). "Small Area Estimates of the Child Population and Poverty in School Districts Using Dirichlet-Multinomial Models," 2019 Proceedings of the American Statistical Association, Section on Survey Research Methods, American Statistical Association, Alexandria, VA, 3150-3152.
- Bell, W. R., Chung, H. C., Datta, G. S., and Franco, C. (2019). "Measurement Error in Small Area Estimation: Functional vs. Structural vs. Naïve Models," *Survey Methodology*, *45*, 61-80.
- Chakraborty, A., Datta, G.S., and Mandal, A. (2019). "Robust Hierarchical Bayes Small Area Estimation for Nested Error Regression Model," *International Statistical Review*, 87, S1, S158–S176, doi:10.1111/insr.12283.
- Chung, H., Datta, G., and Maples, J. (2019). "Estimation of Median Incomes of the American States: Bayesian Estimation of Means of Subpopulations," *Opportunities and Challenges in Development*, Simanti Bandyopadhyay and Mousumi Datta (ed.), New York: Springer, 505-518.
- Franco, C., Little, R. J. A., Louis, T. A., and Slud, E. V. (2019). "Comparative Study of Confidence Intervals for Proportions in Complex Surveys," *Journal of Survey Statistics and Methodology*, 7, 3, 334-364.
- Datta, G.S., Rao, J.N.K., Torabi, M., and Liu, B. (2018). "Small Area Estimation with Multiple Covariates Measured with Errors: A Nested Error Linear Regression Approach of Combining Two Surveys," *Journal of Multivariate Analysis, 167*, 49-59.
- Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017). "Multivariate Fay-Herriot Bayesian Estimation of Small Area Means Under Functional Measurement Error," *Journal of the Royal Statistical Society--Series A*, 180(4), 1191-1209.

- Janicki, R. and Vesper, A. (2017). "Benchmarking Techniques for Reconciling Small Area Models at Distinct Geographic Levels," *Statistical Methods Applications*, DOI: <u>https://doi.org/10.1007/s10260-017-0379-x</u>, 26, 557-581.
- Maples, J. (2017). "Improving Small Area Estimates of Disability: Combining the American Community Survey with the Survey of Income and Program Participation," *Journal of the Royal Statistical Society-Series A*, 180(4), 1211-1227.
- Chakraborty, A., Datta, G.S., and Mandal, A. (2016). "A Two-component Normal Mixture Alternative to the Fay-Herriot Model," Joint issue of Statistics in Transition new series and Survey Methodology, Part II, 17, 67-90.
- Janicki, R (2016). "Estimation of the Difference of Small Area Parameters from Different Time Periods," *Research Report Series* (*Statistics #2016-01*), Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Datta, G.S. and Mandal, A. (2015). "Small Area Estimation with Uncertain Random Effects," *Journal of the American Statistical Association: Theory and Methods*, 110, 1735-1744.
- Franco, C. and Bell, W. R. (2015). "Borrowing Information over Time in Binomial/logit Normal Models for Small Area Estimation," *Joint issue of Statistics in Transition and Survey Methodology*, *16*, *4*, 563-584.
- Bell, W.R., Datta, G.S., and Ghosh, M. (2013). "Benchmarking Small area Estimators," *Biometrika*, 100, 189-202, doi:10.1093/biomet/ass063.
- Franco, C. and Bell, W. R. (2013). "Applying Bivariate/Logit Normal Models to Small Area Estimation," In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 690-702.
- Datta, G., Ghosh, M., Steorts, R., and Maples, J. (2011). "Bayesian Benchmarking with Applications to Small Area Estimation," *TEST, Volume 20, Number 3*, 574-88.
- Janicki, R. (2011). "Selection of Prior Distributions for Multivariate Small Area Models with Application to Small Area Health Insurance Estimates," *JSM Proceedings, Government Statistics Section*. American Statistical Association, Alexandria, VA.
- Maples, J. (2011). "Using Small-Area Models to Improve the Design-Based Estimates of Variance for County Level Poverty Rate Estimates in the American Community Survey," *Research Report Series (Statistics #2011-02)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Slud, E. and Maiti, T. (2011). "Small-Area Estimation Based on Survey Data from Left-Censored Fay-Herriot Model," Journal of Statistical Planning & Inference, 3520-3535.
- Joyce, P. and Malec, D. (2009). "Population Estimation Using Tract Level Geography and Spatial Information," *Research Report* Series (Statistics #2009-3), Statistical Research Division, U.S. Census Bureau, Washington, D.C.
- Malec, D. and Maples, J. (2008). "Small Area Random Effects Models for Capture/Recapture Methods with Applications to Estimating Coverage Error in the U.S. Decennial Census," *Statistics in Medicine*, *27*, 4038-4056.
- Malec, D. and Müller, P. (2008). "A Bayesian Semi-Parametric Model for Small Area Estimation," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* (eds. S. Ghoshal and B. Clarke), Institute of Mathematical Statistics, 223-236.
- Huang, E., Malec, D., Maples J., and Weidman, L. (2007). "American Community Survey (ACS) Variance Reduction of Small Areas via Coverage Adjustment Using an Administrative Records Match," *Proceedings of the 2006 Joint Statistical Meetings*, American Statistical Association, Alexandria, VA, 3150-3152.
- Maples, J. and Bell, W. (2007). "Small Area Estimation of School District Child Population and Poverty: Studying Use of IRS Income Tax Data," *Research Report Series (Statistics #2007-11)*, Statistical Research Division, U.S. Census Bureau, Washington, D.C.
- Slud, E. and Maiti, T. (2006). "Mean-Squared Error Estimation in Transformed Fay-Herriot Models," *Journal of the Royal Statistical Society-Series B*, 239-257.
- Malec, D. (2005). "Small Area Estimation from the American Community Survey Using a Hierarchical Logistic Model of Persons and Housing Units," *Journal of Official Statistics*, 21 (3), 411-432.

Contact: Jerry Maples, Carolina Franco, Ryan Janicki, Gauri Datta, Kyle Irimata, Bill Bell (ADRM), Eric Slud

**Funding Sources for FY 2021:** 0331 – Working Capital Fund / General Research Project Various Decennial, Demographic, and Economic Projects

#### **Sampling Estimation & Survey Inference**

**Motivation:** Survey sampling helps the Census Bureau provide timely and cost efficient estimates of population characteristics. Demographic sample surveys estimate characteristics of people or households such as employment, income, poverty, health, insurance coverage, educational attainment, or crime victimization. Economic sample surveys estimate characteristics of businesses such as payroll, number of employees, production, sales, revenue, or inventory. Survey sampling helps the Census Bureau assess the quality of each decennial census. Estimates are produced by use of design-based estimation techniques or model-based estimation techniques. Methods and topics across the three program areas (Demographic, Economic, and Decennial) include: sample design, estimation and use of auxiliary information (e.g., sampling frame and administrative records), weighting methodology, adjustments for non-response, proper use of population estimates as weighting controls, variance estimation, effects of imputation on variances, coverage measurement sampling and estimation, coverage measurement evaluation, evaluation of census operations, uses of administrative records in census operations, improvement in census processing, and analyses that aid in increasing census response.

#### **Research Problems:**

- How to design and analyze sample surveys from "frames" determined by non-probabilistically sampled observational data to achieve representative population coverage. To make census data products based jointly on administrative and survey data fully representative of the general population, as our current surveys are, new sampling designs and analysis methods will have to be developed.
- How can inclusion in observational or administrative lists be modeled jointly with indicator and mode of survey response, so that traditional survey methods can be extended to merged survey and non-survey data?
- Can non-traditional design methods such as adaptive sampling be used to improve estimation for rare characteristics and populations?
- How can time series and spatial methods be used to improve ACS estimates or explain patterns in the data?
- Can generalized weighting methods be formulated and solved as optimization problems to avoid the ambiguities resulting from multiple weighting step and to explicitly allow inexact calibration?
- What models can aid in assessing the combined effect of all the sources of sampling and nonsampling error, including frame coverage errors and measurement errors, on sample survey estimates?
- What experiments and analyses can inform the development of outreach methods to enhance census response?
- Can unduplication and matching errors be accounted for in modeling frame coverage in censuses and sample surveys?
- How can small-area or other model-based methods be used to improve interval estimates in sample surveys, to design survey collection methods with lowered costs, or to improve Census Bureau imputation methods?
- Can classical methods in nonparametrics (e.g., using ranks) improve estimates from sample surveys?
- How can we measure and present uncertainty in rankings of units based on sample survey estimates?
- Can Big Data improve results from censuses and sample surveys?
- How to develop and use bootstrap methods for expressing uncertainty in estimates from probability sampling?

#### **Current Subprojects:**

- Optimization-based (single-stage) approaches to Weight-adjustment for Probability and Nonprobability Samples (Slud, Morris)
- The Ranking Project: Methodology Development and Evaluation (Wright, Klein/FDA, Wieczorek/Colby College, Yau)
- Optimal Sample Allocation and Apportionment (Wright)
- Optimal stratification in economic surveys, using multiple measures of size and multiple survey outcomes (Slud, Joyce)
- Machine Learning projects related to non-response segmentation Mindsets for decennial outreach (Mulry, Morris, Scheid/DSSD), or to Frames (Weinberg, Slud)
- Methods of estimating variances for survey estimates combining model- and design-based estimates, and simulation studies of bias when the design-based methods include Replication Methods in domains with small sample-size (Slud, Trudell)
- Analyses supporting improvement of household rosters for census nonresponders that are projected to be occupied and to have high quality administrative records. (Mulry).

#### **Potential Applications:**

- Improve estimates and reduce costs for household surveys by introducing new design and estimation methods.
- Produce improved ACS small area estimates thorough the use of time series and spatial methods, where those methods improve upon small area methods using covariates recoded from temporal and spatial information.
- Streamline documentation and make weighting methodology more transparent by applying the same nonresponse and

calibration weighting adjustment software across different surveys.

- New procedures for adjusting weights or reported values in the monthly trade surveys and surveys of government employment, based on statistical identification of outliers and influential values, to improve accuracy of estimation monthly level and of month-to-month change.
- Provide a synthesis of the effect of nonsampling errors on estimates of net census coverage error, erroneous enumerations, and omissions and identify the types of nonsampling errors that have the greatest effects. Employ administrative records to improve the estimates of census coverage error.
- Measure and report uncertainty in rankings in household and economic sample surveys.
- Develop bootstrap methods for expressing uncertainty as an alternative source of published variance estimates and as a check on existing methods of producing variances in Census Bureau sample surveys.

#### Accomplishments (October 2018-September 2020):

- Contributed to team development of methods for producing differentially private decennial census tabulations conforming to legally mandated error-free disclosure of block-level population totals under *Public Law 94* as well as to *Title 13* requirements for nondisclosure of individual-level data.
- Developed novel optimization-based weighting adjustment methods based on partially missing data, along with diagnostics based on cross-classified post-stratification variables
- Demonstrated the potential for a market segmentation from an external source to improve self-response propensity models using data from the 2010 Census and the American Community Survey.
- Demonstrated that market segmentation from an external source aid in providing useful information about problems in the Census enumeration of young children.
- Established theoretical limitations on consistent estimation of variance component parameters from informatively sampled complex survey data based only on single-inclusion weights.
- Developed a simple and novel measure of uncertainty for an estimated ranking with theory, using American Community Survey travel time to work data, and with a visualization.
- Extended the current *equal proportions* methodology by appealing to probability sampling results.
- Developed a general exact optimal sample allocation algorithm with bounded cost and bounded stratum sample sizes.

#### Short-Term Activities (FY 2021):

- Extend Machine Learning approaches to non-response segmentation and frame changes.
- Develop optimal stratification in economic surveys, using multiple measures of size and multiple survey outcomes.
- Document biases of SDR design-based variance estimates for survey-weighted totals in small domains, and what survey design and attribute features they depend on.
- Continue research into post-stratified weight adjustment methodology and assessment of weights, with application to low-response probability surveys and non-probability data collection as in the Tracking Survey.
- Extend research into stratification methodology for economic surveys based on multiple MOS variables and multiple outcomes.
- Continue research into alternative techniques for statistical nondisclosure control motivated by randomize-response techniques
- Improve methodology for measuring uncertainty in rankings.
- Extend methodology for exact optimal sample allocation and apportionment.

#### Longer-Term Activities (beyond FY 2021):

- Extension of Census Matching capability to non-PIK persons using Administrative Records, Duplicate Status and Post-Enumeration Survey data for evaluation of Matching quality.
- Develop software that is re-usable and easily implementable for small area prediction within language minority groups in connection with the determinations of ballot language assistance by jurisdiction and American Indian Area under Section 203 of the Voting Rights Act.
- Further investigate the statistical implications and assumptions of formal privacy (e.g. differential privacy) methods in order to understand how the methods may impact the use of data products and to develop estimates of variability of released data that has been privatized by noise infusion.
- Develop statistical methods and theory related to the use of differential privacy to release data from unequal probability sampling

surveys. A specific focus of this research would be on how to account for the sampling probabilities/weights in the planning of the privacy budget.

- Develop probability sampling methods targeted to the complement of an administrative records database within a survey frame such as the MAF; this research will require combining statistical models for joint dependence of administrative records and survey or census response, to be incorporated into new response propensity models in terms of which the survey data can be analyzed.
- Develop spatial models and associated small area estimation techniques in terms of Generalized Linear Mixed Models (GLMMs) with covariates recoded to incorporate local spatial geographic/demographic/economic effects, and compare the performance of these models with Bayes-hierarchical models currently being developed elsewhere at the Census Bureau using American Community Survey data. Such GLMM spatial models may also be applicable to the evaluation of canvassing and address status changes in the MAF.

#### Selected Publications:

- Trudell, T., Dong, K., Slud, E., and Cheng, Y. (In Press). "Computing Replicated Variance for Stratified Systematic Sampling," Proceedings of the Survey Research Methods Section of the American Statistical Association.
- Wright, T. (In Press). "A General Exact Optimal Sample Allocation Algorithm: With Bounded Cost and Bounded Sample Sizes," *Statistics and Probability Letters.*
- Klein M., Wright, T., and Wieczorek, J. (2020). "A Joint Confidence Region for an Overall Ranking of Population," *Journal of the Royal Statistical Society, Series C, 69, Part 3,* 589-606.
- Mulry, M., Bates, N., and Virgile, M. (2020). "Viewing Participation in Censuses and Surveys through the Lens of Lifestyle Segments." *Journal of Survey Statistics and Methodology*. doi: 10.1093/jssam/smaa006
- Franco, C., Little, R., Louis, T., and Slud, E. (2019). "Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys," *Journal of Survey Statistics and Methodology*, 7, 334-364.
- Slud, E. and <u>Thibaudeau</u>, Y. (2019). "Multi-Outcome Longitudinal Small Area Estimation, A Case Study," *Statistical Theory and Related Fields. Special Issue on Small Area Estimation*, *3*, 136-149.

Wright, T., Klein, M., and Wieczorek, J. (2019). "A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals," *The American Statistician, Vol 73, No 2*, 165-178.

- Chai, J. and Nayak, T. (2018). "A Criterion for Privacy Protection in Data Collection and its Attainment via Randomized Response Procedures," Electronic Journal of Statistics 12 (2), 4264-4287.
- de Oliveira, V., Wang, B., and Slud, E. (2018). "Spatial Modeling of Rainfall Accumulated over Short Periods of Time," *Journal of Multivariate Analysis*, 166, 129-149.
- Dong, K., Trudell, T., Slud, E., and Cheng, Y. (2018). "Understanding Variance Estimator Bias in Stratified Two-Stage Sampling," Proceedings of the Survey Research Methods Section of the American Statistical Association.
- Klein, M., Wright, T., and Wieczorek, J. (2018). "A Simple Joint Confidence Region for A Ranking of K Populations: Application to American Community Survey's Travel Time to Work Data," *Research Report Series (Statistics #2018-04)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Lu, B. and Ashmead, R. (2018). "Propensity Score Matching Analysis for Causal Effects with MNAR Covariates," *Statistica Sinica*, 28, 2005-2025.
- Mulry, M.H, Kaputa, S., and Thompson, K. (2018). "Initial M-estimation Parameter Settings for Detection and Treatment of Influential Values," *Journal of Official Statistics*, 34(2). 483–501. http://dx.doi.org/10.2478/JOS-2018-0022
- Nayak, T., Zhang, C., and You, J. (2018). "Measuring Identification Risk in Microdata Release and Its Control by Post-randomisation," *International Statistical Review*, 86 (2), 300-321.
- Slud, E., Vonta, I., and Kagan, A. (2018). "Combining Estimators of a Common Parameter across Samples," *Statistical Theory and Related Fields*, 2, 158-171.
- Wright, T. (2018). "No Calculation When Observation Can Be Made," in A.K. Chattopadhyay and G. Chattopadhyay (Eds), *Statistics and Its Applications*, Springer Singapore, 139-154.
- Ashmead, R., Slud, E., and Hughes, T. (2017). "Adaptive Intervention Methodology for Reduction of Respondent Contact Burden in the American Community Survey," *Journal of Official Statistics*, 33(4), 901-919.
- Ashmead, R. and Slud, E. (2017). "Small Area Model Diagnostics and Validation with Applications to the Voting Rights Act Section 203," *Proceedings of Survey Research Methods Section*, American Statistical Association, Alexandria, VA.
- Mulry, M.H. and Keller, A. (2017). "Comparison of 2010 Census Nonresponse Follow-up Proxy Responses with Administrative Records Using Census Coverage Measurement Results," *Journal of Official Statistics*, 33(2), 455–475. DOI: https://doi.org/10.1515/jos-2017-0022
- Mulry, M.H., Nichols, E. M., and Hunter Childs, J. (2017). "Using Administrative Records Data at the U.S. Census Bureau: Lessons Learned from Two Research Projects Evaluating Survey Data." In Biemer, P.P., Eckman, S., Edwards, B., Lyberg, L., Tucker, C., de Leeuw, E., Kreuter, F., and West, B.T. *Total Survey Error in Practice*. Wiley. New York. 467-473.
- Slud, E. and Ashmead, R. (2017). "Hybrid BRR and Parametric-Bootstrap Variance Estimates for Small Domains in Large Surveys," *Proceedings of Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

- Thibaudeau, Y., Slud, E., and Gottschalck, A. (2017). "Modeling Log-linear Conditional Probabilities for Estimation in Surveys," *Annals of Applied Statistics*, 11 (2), 680-697.
- Wieczorek, J. (2017). "Ranking Project: The Ranking Project: Visualizations for Comparing Populations," R package version 0.1.1. URL: https://cran.r-project.org/package=RankingProject.
- Wright, T. (2017). "Exact Optimal Sample Allocation: More Efficient Than Neyman," Statistics and Probability Letters, 129, 50-57.
- Mulry, M. H., Nichols, E. M., and Childs, J. Hunter (2016). "A Case Study of Error in Survey Reports of Move Month Using the U.S. Postal Service Change of Address Records," Survey Methods: Insights from the Field. Retrieved from http://surveyinsights.org/?p=7794
- Mulry, M.H., Oliver, B., Kaputa, S., and Thompson, K. J. (2016). "Cautionary Note on Clark Winsorization." Survey Methodology 42 (2), 297-305. http://www.statcan.gc.ca/pub/12-001-x/2016002/article/14676-eng.pdf
- Nayak, T. and Adeshiyan, S. (2016). "On Invariant Post-randomization for Statistical Disclosure Control," International Statistical Review, 84 (1), 26-42.
- Nayak, T., Adeshiyan, S. and Zhang, C. (2016). "A Concise Theory of Randomized Response Techniques for Privacy and Confidentiality Protection," Handbook of Statistics, 34, 273-286.
- Wright, T. (2016). "Two Optimal Exact Sample Allocation Algorithms: Sampling Variance Decomposition Is Key," Research Report Series (Statistics #2016-03), Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Nagaraja, C. and McElroy, T. (2015). "On the Interpretation of Multi-Year Estimates of the American Community Survey as Period Estimates." Published online, *Journal of the International Association of Official Statistics*.
- Slud, Eric. (2015). "Impact of Mode-based Imputation on ACS Estimates," American Community Survey Research and Evaluation Memorandum, #ACS-RER-07.
- Franco, C., Little, R., Louis, T., and Slud, E. (2014). "Coverage Properties of Confidence Intervals for Proportions in Complex Sample Surveys," *Proceedings of Survey Research Methods Section*, American Statistical Association, Alexandria, VA.
- Griffin, D., Slud, E., and Erdman, C. (2014). "Reducing Respondent Burden in the American Community Survey's Computer Assisted Personal Visit Interviewing Operation - Phase 3 Results," ACS Research and Evaluation Memorandum #ACS 14-RER-28.
- Hogan, H. and Mulry, M. H. (2014). "Assessing Accuracy of Postcensal Estimates: Statistical Properties of Different Measures," in N. Hogue (Ed.), *Emerging Techniques in Applied Demography*. Springer. New York.
- Hunley, Pat. (2014). "Proof of Equivalence of Webster's Method and Willcox's Method of Major Fractions," Research Report Series (Statistics #2014-04), Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Joyce, P., Malec, D., Little, R., Gilary, A., Navarro, A., and Asiala, M. (2014). "Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations," *Journal of American Statistical Association*, 109 (505), 36-47.
- Mulry, M. H. (2014). "Measuring Undercounts in Hard-to-Survey Groups," in R. Tourangeau, N. Bates, B. Edwards, T. Johnson, and K. Wolter (Eds.), *Hard-to-Survey Populations*. Cambridge University Press, Cambridge, England.
- Mulry, M. H., Oliver, B. E., and Kaputa, S. J. (2014) "Detecting and Treating Verified Influential Values in a Monthly Retail Trade Survey." *Journal of Official Statistics*, 30(4), 1–28.
- Shao, J., Slud, E., Cheng, Y., Wang, S., and Hogue, C. (2014). "Theoretical and Empirical Properties of Model Assisted Decision-Based Regression Estimators," Survey Methodology 40(1), 81-104.
- Tang, M., Slud, E., and Pfeiffer, R. (2014). "Goodness of Fit Tests for Linear Mixed Models," *Journal of Multivariate Analysis*, 130, 176-193.
- Wright, T. (2014). "A Simple Method of Exact Optimal Sample Allocation under Stratification with Any Mixed Constraint Patterns," *Research Report Series (Statistics #2014-07)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Wright, T. (2014). "Lagrange's Identity and Congressional Apportionment," The American Mathematical Monthly, 121, 523-528.
- Slud, E., Grieves, C., and Rottach, R. (2013). "Single Stage Generalized Raking Weight Adjustment in the Current Population Survey," Proceedings of Survey Research Methods Section, American Statistical Association, Alexandria, VA.
- Wright, T. (2013). "A Visual Proof, a Test, and an Extension of a Simple Tool for Comparing Competing Estimates," *Research Report Series (Statistics #2013-05)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Wright, T., Klein, M., and Wieczorek, J. (2013). "An Overview of Some Concepts for Potential Use in Ranking Populations Based on Sample Survey Data," 2013 Proceedings of the World Congress of Statistics (Hong Kong), International Statistical Institute.
- Ikeda, M., Tsay, J., and Weidman, L. (2012). "Exploratory Analysis of the Differences in American Community Survey Respondent Characteristics between the Mandatory and Voluntary Response Methods," *Research Report Series (Statistics* #2012-01), Center for Statistical Research & Methodology, U.S. Census Bureau, Wash. D.C.
- Wright, T. (2012). "The Equivalence of Neyman Optimum Allocation for Sampling and Equal Proportions for Apportioning the U.S. House of Representatives," *The American Statistician*, 66 (4), 217-224.
- Klein, M. and Wright, T. (2011). "Ranking Procedures for Several Normal Populations: An Empirical Investigation," International Journal of Statistical Sciences, Volume 11 (P.C. Mahalanobis Memorial Special Issue), 37-58.
- Slud, E. and Thibaudeau, Y. (2010). "Simultaneous Calibration and Nonresponse Adjustment," Research Report Series

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-2 Filed 04/20/21 Page 19 of 31

(Statistics#2010-03), Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Contact: Eric Slud, Mary Mulry, Michael Ikeda, Patrick Joyce, Tapan Nayak, Edward H. Porter, Tommy Wright

Funding Sources for FY 2021:0331 – Working Capital Fund / General Research ProjectVarious Decennial, Demographic, and Economic Projects

#### **Time Series & Seasonal Adjustment**

**Motivation:** Seasonal adjustment is vital to the effective presentation of data collected from monthly and quarterly economic sample surveys by the Census Bureau and by other statistical agencies around the world. As the developer of the *X-13ARIMA-SEATS Seasonal Adjustment Program*, which has become a world standard, it is important for the Census Bureau to maintain an ongoing program of research related to seasonal adjustment methods and diagnostics, in order to keep *X-13ARIMA-SEATS* up-to-date and to improve how seasonal adjustment is done at the Census Bureau.

#### **Research Problems:**

- All contemporary seasonal adjustment programs of interest depend heavily on time series models for trading day and calendar effect estimation, for modeling abrupt changes in the trend, for providing required forecasts, and, in some cases, for the seasonal adjustment calculations. Better methods are needed for automatic model selection, for detection of inadequate models, and for assessing the uncertainty in modeling results due to model selection, outlier identification and non-normality. Also, new models are needed for complex holiday and calendar effects.
- Diagnostics of seasonality must address differing sampling frequencies (monthly versus quarterly) and multiple forms of seasonality (cycles of annual versus weekly period), and must distinguish between raw and seasonally adjusted data.
- Multivariate modeling can not only provide increased precision of seasonal adjustments, but can also assist with series that have a low signal content. Moreover, multivariate techniques expand the class of univariate models, allowing the modeling of seasonal heteroscedasticity. This motivates the need to develop a viable multivariate seasonal adjustment methodology that can handle modeling, fitting, and seasonal adjustment of a large number of series.
- Time series data are being measured at higher sampling rates or over geographical regions, requiring new seasonal adjustment methods for high frequency/space-time data.
- Many published time series arise from sample surveys, and are subject to sampling error. Methodology and algorithms are needed to incorporate sampling error components into the existing seasonal adjustment framework.

#### **Current Subprojects:**

- Seasonal Adjustment (McElroy/ADRM, Livsey, Pang, Roy)
- Time Series Analysis (McElroy/ADRM, Livsey, Pang, Roy, Trimbur)

#### **Potential Applications**

• Applications encompass the Decennial, Demographic, and Economic areas.

#### Accomplishments (October 2018-September 2020):

- Developed and implemented new algorithms for ragged edge missing value imputation, and ad hoc filtering of multivariate time series.
- Implemented and tested autoregressive diagnostics for seasonality.
- Refined a benchmarking method to remove seasonality from indirect seasonal adjustments.
- Added new models with stable parameterizations to multivariate time series software.
- Studied an EM approach to modeling multivariate time series.
- Studied outlier processes, allowing for a new approach to extreme-value adjustment of seasonal time series.
- Developed methods and formulas for quadratic filtering and forecasting of time series.

#### Short-Term Activities (FY 2021):

- Continue developing diagnostics for seasonality by refining the AR diagnostic and examining forecast error and partial autocorrelation.
- Continue the study of weekly and daily time series, including the facets of modeling, fitting, computation, separation of low-frequency signals, identification of holiday effects, attenuating of extremes, and applications to change of support problems.
- Develop nonlinear filtering and prediction methods based on autocumulants, with applications to seasonal adjustment in the presence of extremes.
- Develop improved automatic model identification methods.
- Develop extensions to maximum entropy extreme-value framework, allowing for more general types of outliers.
- Generate an R package for Ecce Signum, and disseminate X-13 R Story.
- Continue examining methods for estimating trading day regressors with time-varying coefficients, and determine which Census Bureau series are amenable to moving trading day adjustment.
- Study the impact of sampling error on seasonal adjustment.

#### Longer-Term Activities (beyond FY 2021):

- Further develop methods for constrained signal extraction, appropriate for multivariate data subject to accounting relations.
- Continue investigation of Seasonal Vector Form, allowing for more exotic seasonal models, and develop the corresponding seasonal adjustment methods.
- Expand research on multivariate seasonal adjustment in order to address the facets of co-integration, batch identification, modeling, estimation, and algorithms.
- Improve the speed and stability of likelihood optimization in X-13ARIMA-SEATS.
- Investigate the properties and applications of both integer time series and network time series models.
- Develop and disseminate software to implement state space models, with the intention of treating sampling error and stochastic trading day.
- Develop estimators for the duration of a moving holiday effect.
- Continue investigation of cycles, band-pass filters, and signal extraction machinery for a broad array of signals.

#### Selected Publications:

- Baker, S., McElroy, T.S., and Sheng, X. (2020). "Expectation Formation Following Large and Unpredictable Shocks," *Review of Economics and Statistics*, 14, 112-130.
- McElroy, T.S. and Politis, D.N. (2020). Time Series: a First Course with Bootstrap Starter. New York: Chapman Hall.
- McElroy, T.S. and Wildi, M. (2020). "Multivariate Direct Filter Analysis for Real-Time Signal Extraction Problems," *Econometrics and Statistics*, 14, 112-130.
- Hyatt, H. and McElroy, T.S. (2019). "Labor Reallocation, Employment, and Earnings: Vector Autoregression Evidence," *LABOUR*, 33(4), 463-487.
- McElroy, T.S. and Jach, A. (2019). "Testing Collinearity of Vector Time Series," The Econometrics Journal, 22(2), 97-116.
- McElroy, T. S. and Jach, A. (2019). "Subsampling Inference for the Autocorrelations of GARCH Processes," Published online, *Journal of Financial Econometrics*, 17(3), 495-515.
- McElroy, T. S., Pang, O., and Sheldon, G. (2019). "Custom Epoch Estimation for Surveys," Published online, *Journal of Applied Statistics*, 46, 638-663.
- McElroy, T.S. and Penny, R. (2019). "Maximum Entropy Extreme-Value Seasonal Adjustment," Australian New Zealand Journal of Statistics, 61(2), 152-174.
- Roy, A., McElroy, T. S., and Linton, P. (2019). "Estimation of Causal Invertible VARMA Models," *Statistica Sinica*, 29(1), 455-478.
- Wildi, M. and McElroy, T. S. (2019). "The Trilemma between Accuracy, Timeliness, and Smoothness in Real-Time Signal Extraction," *International Journal of Forecasting*, *35*, 1072-1084
- Findley, D.F. and McElroy, T. S. (2018). "Background and Perspectives for ARIMA Model-Based Seasonal Adjustment," *Research Report Series (Statistics #2018-07)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Lin, W., Huang, J., and McElroy, T. S. (2018). "Time Series Seasonal Adjustment Using Regularized Singular Value Decomposition," Published online, *Journal of Business and Economics Statistics*.
- Livsey, J., Lund, R., Kechagias, S., and Pipiras, V. (2018). "Multivariate Integer-valued Time Series with Flexible Autocovariances and Their Application to Major Hurricane Counts," *Annals of Applied Statistics, 12(1):* 408-431.
- McElroy, T. S. (2018). "Recursive Computation for Block Nested Covariance Matrices," *Journal of Time Series Analysis, 39 (3),* 299-312.
- McElroy, T. S. (2018). "Seasonal Adjustment Subject to Accounting Constraints," Statistica Neerlandica, 72, 574-589.
- McElroy, T.S., Monsell B. C., and Hutchinson, R. (2018). "Modeling of Holiday Effects and Seasonality in Daily Time Series," *Research Report Series (Statistics 2018-01)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- McElroy, T.S. and Roy, A. (2018). "Model Identification via Total Frobenius Norm of Multivariate Spectra," *Research Report Series (Statistics #2018-03)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- McElroy, T. S. and Roy, A. (2018). "The Inverse Kullback Leibler Method for Fitting Vector Moving Averages," *Journal of Time Series Analysis, 39*, 172-191.
- Nagaraja, C. and McElroy, T. S. (2018). "The Multivariate Bullwhip Effect," *European Journal of Operations Research*, 267, 96-106.
- Blakely, C. and McElroy, T. S. (2017). "Signal Extraction Goodness-of-fit Diagnostic Tests under Model Parameter Uncertainty: Formulations and Empirical Evaluation," *Econometric Reviews*, *36* (4), 447-467.
- Findley, D.F., Lytras, D. P., and McElroy, T. S. (2017). "Detecting Seasonality in Seasonally Adjusted Monthly Time Series," *Research Report Series (Statistics #2017-03)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.

- Holan, S., McElroy, T. S., and Wu, G. (2017). "The Cepstral Model for Multivariate Time Series: The Vector Exponential Model," *Statistica Sinica* 27, 23-42.
- McElroy, T. S. (2017). "Computation of Vector ARMA Autocovariances," Statistics and Probability Letters, 124, 92-96.
- McElroy, T. S. (2017). "Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy," *Journal of Business* and Economics Statistics, 35 (4), 511-525.
- McElroy, T. S. and McCracken, M. (2017). "Multi-Step Ahead Forecasting of Vector Time Series," *Econometric Reviews*, 36 (5), 495-513.
- McElroy, T.S. and Monsell, B. C (2017). "Issues Related to the Modeling and Adjustment of High Frequency Time Series," *Research Report Series (Statistics #2017-08)*, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Sanyal, A., Mitra, P., McElroy, T.S., and Roy, A. (2017). "Holiday Effects in Indian Manufacturing Series," *Research Report* Series (Statistics #2017-04), Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, D.C.
- Trimbur, T. and McElroy, T. S. (2017). "Signal Extraction for Nonstationary Time Series with Diverse Sampling Rules," *Journal* of Time Series Econometrics, 9 (1).
- Janicki, R. and McElroy, T. (2016). "Hermite Expansion and Estimation of Monotonic Transformations of Gaussian Data," *Journal* of Nonparametric Statistics, 28(1), 207-234.
- McElroy, T. S. (2016). "Non-nested Model Comparisons for Time Series," Biometrika, 103, 905-914.
- McElroy, T. (2016). "On the Measurement and Treatment of Extremes in Time Series," Extremes, 19(3), 467-490.
- McElroy, T. and Nagaraja, C. (2016). "Tail Index Estimation with a Fixed Tuning Parameter Fraction," *Journal of Statistical Planning and Inference*, 170, 27-45.
- Trimbur, T. and McElroy, T. (2016). "Signal Extraction for Nonstationary Time Series with Diverse Sampling Rules," Published online, *Journal of Time Series Econometrics*.
- Wildi, M. and McElroy, T. (2016). "Optimal Real-Time Filters for Linear Prediction Problems," *Journal of Time Series Econometrics*, 8(2), 155-192.
- Lund, R., Holan, S., and Livsey, J. (2015). "Long Memory Discrete-Valued Time Series." Forthcoming, *Handbook of Discrete-Valued Time Series*. Eds R. Davis, S. Holan, R. Lund, N. Ravishanker. CRC Press.
- Lund, R. and Livsey, J. (2015). "Renewal Based Count Time Series." Forthcoming, *Handbook of Discrete-Valued Time Series*. Eds R. Davis, S. Holan, R. Lund, N. Ravishanker. CRC Press.
- McElroy, T. (2015). "When are Direct Multi-Step and Iterative Forecasts Identical?" Journal of Forecasting, 34, 315-336.
- McElroy, T. and Findley, D. (2015). "Fitting Constrained Vector Autoregression Models," in *Empirical Economic and Financial Research*.
- McElroy, T. and Monsell, B. (2015). "Model Estimation, Prediction, and Signal Extraction for Nonstationary Stock and Flow Time Series Observed at Mixed Frequencies." *Journal of the American Statistical Association (Theory and Methods), 110,* 1284-1303.
- McElroy, T. and Pang, O. (2015). "The Algebraic Structure of Transformed Time Series," in *Empirical Economic and Financial Research*.
- McElroy, T. and Trimbur, T. (2015). "Signal Extraction for Nonstationary Multivariate Time Series with Illustrations for Trend Inflation." *Journal of Time Series Analysis 36*, 209--227. Also in "Finance and Economics Discussion Series," Federal Reserve Board. 2012-45. http://www.federalreserve.gov/pubs/feds/2012/201245/201245abs.html
- McElroy, T. and Holan, S. (2014). "Asymptotic Theory of Cepstral Random Fields," Annals of Statistics, 42, 64-86.
- McElroy, T. and Maravall, A. (2014). "Optimal Signal Extraction with Correlated Components," *Journal of Time Series Econometrics*, 6, 237--273.
- McElroy, T. and Monsell, B. (2014). "The Multiple Testing Problem for Box-Pierce Statistics," *Electronic Journal of Statistics*, 8, 497-522.
- McElroy, T. and Politis, D. (2014). "Spectral Density and Spectral Distribution Inference for Long Memory Time Series via Fixed-b Asymptotics," *Journal of Econometrics*, 182, 211-225.
- Monsell, B. C. (2014) "The Effect of Forecasting on X-11 Adjustment Filters," 2014 Proceedings American Statistical Association [CD-ROM]: Alexandria, VA.
- Roy, A., McElroy, T., and Linton, P. (2014). "Estimation of Causal Invertible VARMA Models," *Cornell University Library*, http://arxiv.org/pdf/1406.4584.pdf.
- Findley, D. F. (2013). "Model-Based Seasonal Adjustment Made Concrete with the First Order Seasonal Autoregressive Model," Center for Statistical Research & Methodology, *Research Report Series (Statistics #2013-04)*. U.S. Census Bureau, Washington, D.C.
- McElroy, T. (2013). "Forecasting CARIMA Processes with Applications to Signal Extraction," Annals of the Institute of Statistical Mathematics, 65, 439-456.
- McElroy, T. and Politis, D. (2013). "Distribution Theory for the Studentized Mean for Long, Short, and Negative Memory Time Series," *Journal of Econometrics*, 177, 60-74.
- McElroy, T. and Wildi, M. (2013). "Multi-Step Ahead Estimation of Time Series Models," *International Journal of Forecasting* 29, 378-394.

- Monsell, B. C. and Blakely, C. (2013). "X-13ARIMA-SEATS and iMetrica," 2013 Proceedings of the World Congress of Statistics (Hong Kong), International Statistical Institute.
- Alexandrov, T., Bianconcini, S., Dagum, E., Maass, P., and McElroy, T. (2012). "The Review of Some Modern Approaches to the Problem of Trend Extraction," *Econometric Reviews*, 31, 593-624.
- Bell, W., Holan, S., and McElroy, T. (2012). Economic Time Series: Modeling and Seasonality. New York: Chapman Hall.
- Blakely, C. (2012). "Extracting Intrinsic Modes in Stationary and Nonstationary Time Series Using Reproducing Kernels and Quadratic Programming," *International Journal of Computational Methods, Vol. 8, No. 3.*
- Findley, D. F., Monsell, B. C., and Hou, C.-T. (2012). "Stock Series Holiday Regressors Generated from Flow Series Holiday Regressors," *Taiwan Economic Forecast and Policy*.
- Holan, S. and McElroy, T. (2012). "On the Seasonal Adjustment of Long Memory Time Series," in *Economic Time Series: Modeling and Seasonality*. Chapman-Hall.
- Jach, A., McElroy, T., and Politis, D. (2012). "Subsampling Inference for the Mean of Heavy-tailed Long Memory Time Series," *Journal of Time Series Analysis*, 33, 96-111.
- McElroy, T. (2012). "The Perils of Inferring Serial Dependence from Sample Autocorrelation of Moving Average Series," Statistics and Probability Letters, 82, 1632-1636.
- McElroy, T. (2012). "An Alternative Model-based Seasonal Adjustment that Reduces Over-Adjustment," *Taiwan Economic Forecast and Policy* 43, 33-70.
- McElroy, T. and Holan, S. (2012). "A Conversation with David Findley," Statistical Science, 27, 594-606.
- McElroy, T. and Holan, S. (2012). "On the Computation of Autocovariances for Generalized Geganbauer Processes," *Statistica Sinica* 22, 1661-1687.
- McElroy, T. and Holan, S. (2012). "The Error in Business Cycle Estimates Obtained from Seasonally Adjusted Data," in Economic Time Series: Modeling and Seasonality. Chapman-Hall.
- McElroy, T. and Jach, A. (2012). "Subsampling Inference for the Autocovariances of Heavy-tailed Long-memory Time Series," *Journal of Time Series Analysis*, 33, 935-953.
- McElroy, T. and Jach, A. (2012). "Tail Index Estimation in the Presence of Long Memory Dynamics," *Computational Statistics and Data Analysis*, 56, 266-282.
- McElroy, T. and Politis, D. (2012). "Fixed-b Asymptotics for the Studentized Mean for Long and Negative Memory Time Series," *Econometric Theory*, 28, 471-481.
- Quenneville, B. and Findley, D. F. (2012). "The Timing and Magnitude Relationships between Month-to-Month Changes and Year-to-Year Changes That Make Comparing Them Difficult," *Taiwan Economic Forecast and Policy*, 43, 119-138.

Contact: Tucker McElroy (R&M), James Livsey, Osbert Pang, Anindya Roy, Bill Bell (R&M), Thomas Trimbur.

Funding Sources for FY 2021: 0331 – Working Capital Fund / General Research Project Economic Projects

#### **Experimentation, Prediction, & Modeling**

**Motivation:** Experiments at the Census Bureau are used to answer many research questions, especially those related to testing, evaluating, and advancing survey sampling methods. A properly designed experiment provides a valid, cost-effective framework that ensures the right type of data are collected as well as sufficient sample sizes and power are attained to address the questions of interest. The use of valid statistical models is vital to both the analysis of results from designed experiments and in characterizing relationships between variables in the vast data sources available to the Census Bureau. Statistical modeling is an essential component for wisely integrating data from previous sources (e.g., censuses, sample surveys, and administrative records) in order to maximize the information that they can provide. In particular, linear mixed effects models are ubiquitous at the Census Bureau through applications of small area estimation. Models can also identify errors in data, e.g. by computing valid tolerance bounds and flagging data outside the bounds for further review.

#### **Research Problems:**

- Investigate established methods and novel extensions to support design (e.g., factorial designs), analysis, and sample size determination for Census Bureau experiments.
- Investigate methodology for experimental designs embedded in sample surveys, including large-scale field experiments embedded in ongoing surveys. This includes design-based and model-based analysis and variance estimation incorporating the sampling design and the experimental design (van den Brakel, Survey Methodology, 2005); factorial designs embedded in sample surveys (van den Brakel, Survey Methodology, 2013), and the estimation of interactions; and testing non-response using embedded experiments.
- Identify and develop statistical models (e.g., loglinear models, mixture models, and mixed-effects models), associated methodologies, and computational tools for problems relevant to the Census Bureau.
- Assess the applicability of *post hoc* methods (e.g., multiple comparisons and tolerance intervals) with future designed experiments and when reviewing previous data analyses.
- Construct rectangular nonparametric tolerance regions for multivariate data. Tolerance regions for multivariate data are usually elliptical in shape, but such regions cannot provide information on individual components of the measurement vector. However, such information can be obtained through rectangular tolerance regions.
- Develop a technique for mis-reporting via the COM-Poisson distribution in order to estimate true counts.
- Develop a disclosure policy motivated by the COM-Poisson and related distributions that allows one to protect individual information reported in two-way and multi-way tables.

#### **Current Subprojects:**

- Developing Flexible Distributions and Statistical Modeling for Count Data Containing Dispersion (Sellers, Morris, Raim).
- Design and Analysis Methods for Experiments (Raim, Mathew, Sellers)

#### **Potential Applications:**

- Modeling can help to characterize relationships between variables measured in censuses, sample surveys, and administrative records and quantify their uncertainty.
- Modeling approaches with administrative records can help enhance the information obtained from various sample surveys.
- Experimental design can help guide and validate testing procedures proposed for censuses and surveys. Sample sizes can be determined to achieve desired power using planned designs and statistical procedures.
- Embedded experiments can be used to evaluate the effectiveness of alternative contact strategies.
- The collection of experimental design procedures currently utilized with the American Community Survey can be expanded.
- Fiducial predictors of random effects can be applied to mixed effects models such as those used in small area estimation.
- Rectangular tolerance regions can be applied to multivariate economic data and aid in the editing process by identifying observations that are outlying in one or more attributes and which subsequently should undergo further review. The importance of ratio edits and multivariate/multiple edits is noted in the work of Thompson and Sigman (*Journal of Official Statistics*, 1999) and de Waal, Pannekoek and Scholtus (Handbook of Statistical Data Editing and Imputation, 2011).
- Principled measures of statistical variability can be provided for constructs like the POP Division's Population Estimates.
- Mis-reporting techniques could be used to assess the amount of mis-reporting in historical Census datasets to aid in model development to estimate true survey count outcomes.

• Statistical disclosure limitation constructs would allow the Census Bureau to release statistical measures associated with a general distributional form while protecting individual privacy. These measures would allow one to estimate the form of multi-way tables of interest while masking the true outcomes.

#### Accomplishments (October 2018-September 2020):

- Completed paper on spatio-temporal change of support modeling in R and released stcos R package.
- Addressed issues with COM-Poisson normalizing constant in the COMPoissonReg R package.
- Completed paper on Conway-Maxwell (COM) multinomial distribution and its use in analyzing clustered multinomial datasets that exhibit over- or under-dispersion.
- Developed and released COMMultReg R package to support COM-multinomial paper.
- Completed paper on continuation-ratio logit modeling for sample size determination and analysis of experiments involving sequences of success/failure trials. Such models support the study of nonresponse probabilities under multiple enumeration attempts to each household.
- Completed paper on comparing pairs of discrete distributions via multinomial outcomes to determine if one is closer to a discrete uniform distribution. This was applied to Census Bureau call volume data to determine if a staggered mailing strategy leads to significantly more uniform call distributions than a simpler strategy where mail is sent to all recipients at once.
- Completed development of a one-step autoregressive model for count data motivated by the COM-Poisson distribution.

#### Short-Term Activities (FY 2021):

- Explore COM-multinomial as a model for missing observations in clustered data under a Bayesian setting.
- Extend work on sample size determination with continuation-ratio logit model to a mixed effects setting.
- Develop a multivariate COM-Poisson distribution model.

#### Longer-Term Activities (beyond FY 2021):

- Develop generalized/flexible spatial and time series models motivated by the COM-Poisson distribution.
- Significant progress has been made recently on randomization-based causal inference for complex experiments; Ding (*Statistical Science*, 2017), Dasgupta, Pillai and Rubin (*Journal of the Royal Statistical Society, Series B*, 2015), Ding and Dasgupta (*Journal of the American Statistical Association*, 2016), Mukerjee, Dasgupta and Rubin (*Journal of the American Statistical Association*, 2016), Mukerjee, Dasgupta and Rubin (*Journal of the American Statistical Association*, 2016), Mukerjee, Dasgupta and Rubin (*Journal of the American Statistical Association*, 2018), Branson and Dasgupta (*International Statistical Review*, 2020). It is proposed to adopt these methodologies for analyzing complex embedded experiments, by taking into account the features of embedded experiments (for example, random interviewer effects and different sampling designs).
- Generalize the Kadane et al. (2006) COM-Poisson motivated data disclosure limitation procedure for one-way tables to handle two-way and multi-way tables. Determine the associated sufficient statistics of the bivariate (or multivariate) COM-Poisson distribution and use them to describe the space of feasible tables that can be used to substitute the true contingency table.
- Consider generalizations of the frequentist and Bayesian approaches to address under-reporting described in Winkelmann (1996), Fader and Hardie (2000), Neubauer and Djuras (2009), and Neubauer et al. (2009) to allow for data dispersion via the COM-Poisson distribution.

#### **Selected Publications:**

- Morris, D.S., Raim, A.M., and Sellers, K.F. (In Press). "A Conway-Maxwell-multinomial Distribution for Flexible Modeling of Clustered Categorical Data," *Journal of Multivariate Analysis*. DOI: https://doi.org/10.1016/j.jmva.2020.104651.
- Raim, A.M., Holan, S.H., Bradley, J.R., and Wikle, C.K. (2020). stcos: "Space-Time Change of Support, version 0.3.0," https://cran.r-project.org/package=stcos.
- Sellers K.F., Peng, S.J., and Arab, A. (2020). "A Flexible Univariate Autoregressive Time-series Model for Dispersed Count Data," *Journal of Time Series Analysis*, 41(3): 436-453.
- Zhu, L., Sellers, K., Morris, D., Shmueli, G., and Davenport, D. (2020). cmpprocess: "Flexible Modeling of Count Processes," version 1.1, https://cran.r-project.org/package=cmpprocess
- Raim, A.M., Holan, S.H., Bradley, J.R., and Wikle, C.K. (2019). "Spatio-Temporal Change of Support Modeling for the American Community Survey with R," URL: https://arxiv.org/abs/1904.12092.
- Sellers, K., Lotze, T., and Raim, A. (2019). COMPoissonReg: "Conway-Maxwell-Poisson Regression, version 0.7.0," https://cran.r-project.org/package=COMPoissonReg
- Sellers, K.F. and Young, D. (2019). "Zero-inflated Sum of Conway-Maxwell-Poissons (ZISCMP) Regression with Application to Shark Distributions," *Journal of Statistical Computation and Simulation*, 89 (9): 1649-1673.
- Sellers, K., Morris, D., Balakrishnan, N., and Davenport, D. (2018). multicmp: "Flexible Modeling of Multivariate Count Data

via the Multivariate Conway-Maxwell-Poisson Distribution," version 1.1, https://cran.r-project.org/package=multicmp

- Morris, D.S., Sellers, K.F., and Menger, A. (2017). "Fitting a Flexible Model for Longitudinal Count Data Using the NLMIXED Procedure," SAS Global Forum Proceedings Paper 202-2017, SAS Institute: Cary, NC.
- Raim, A.M., Holan, S.H., Bradley, J.R., and Wikle, C.K. (2017). "A Model Selection Study for Spatio-Temporal Change of Support," in *Proceedings, Government Statistics Section of the American Statistical Association*, Alexandria, VA: American Statistical Association.
- Sellers, K.F., and Morris, D. (2017). "Under-dispersion Models: Models That Are 'Under The Radar'," *Communications in Statistics Theory and Methods*, 46 (24): 12075-12086.
- Sellers K.F., Morris D.S., Shmueli, G., and Zhu, L. (2017). "Reply: Models for Count Data (A Response to a Letter to the Editor)," *The American Statistician*.
- Young, D.S., Raim, A.M., and Johnson, N.R. (2017). "Zero-inflated Modelling for Characterizing Coverage Errors of Extracts from the U.S. Census Bureau's Master Address File," *Journal of the Royal Statistical Society: Series A*. 180(1):73-97.
- Zhu, L., Sellers, K.F., Morris, D.S., and Shmueli, G. (2017). "Bridging the Gap: A Generalized Stochastic Process for Count Data," *The American Statistician*, 71 (1): 71-80.
- Heim, K. and Raim, A.M. (2016). "Predicting Coverage Error on the Master Address File Using Spatial Modeling Methods at the Block Level," In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association.
- Mathew, T., Menon, S., Perevozskaya, I., and Weerahandi, S. (2016). "Improved Prediction Intervals in Heteroscedastic Mixed-Effects Models," *Statistics & Probability Letters, 114,* 48-53.
- Raim, A.M. (2016). "Informing Maintenance to the U.S. Census Bureau's Master Address File with Statistical Decision Theory," In JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association.
- Sellers, K.F., Morris, D.S., and Balakrishnan, N. (2016). "Bivariate Conway-Maxwell-Poisson Distribution: Formulation, Properties, and Inference," *Journal of Multivariate Analysis*, 150:152-168.
- Sellers, K.F. and Raim, A.M. (2016). "A Flexible Zero-inflated Model to Address Data Dispersion," *Computational Statistics and Data Analysis*, 99: 68-80.
- Raim, A.M. and Gargano, M.N. (2015). "Selection of Predictors to Model Coverage Errors in the Master Address File," *Research Report Series: Statistics #2015-04*, Center for Statistical Research and Methodology, U.S. Census Bureau.
- Young, D. and Mathew, T. (2015). "Ratio Edits Based on Statistical Tolerance Intervals," *Journal of Official Statistics 31*, 77-100.
- Klein, M., Mathew, T., and Sinha, B. K. (2014). "Likelihood Based Inference under Noise Multiplication," *Thailand Statistician*. *12(1)*, pp.1-23. URL: http://www.tci-thaijo.org/index.php/thaistat/article/view/34199/28686
- Young, D.S. (2014). "A Procedure for Approximate Negative Binomial Tolerance Intervals," *Journal of Statistical Computation and Simulation*, 84(2), pp.438-450. URL: http://dx.doi.org/10.1080/00949655.2012.715649
- Gamage, G., Mathew, T., and Weerahandi, S. (2013). "Generalized Prediction Intervals for BLUPs in Mixed Models," *Journal of Multivariate Analysis, 120,* 226-233.
- Mathew, T. and Young, D. S. (2013). "Fiducial-Based Tolerance Intervals for Some Discrete Distributions," Computational Statistics and Data Analysis, 61, 38-49.
- Young, D.S. (2013). "Regression Tolerance Intervals," *Communications in Statistics Simulation and Computation*, 42(9), 2040-2055.

Contact: Andrew Raim, Thomas Mathew, Kimberly Sellers, Darcy Morris

**Funding Sources for FY 2021:** 0331 – Working Capital Fund / General Research Project Various Decennial and Demographic Projects

#### Simulation, Data Science, & Visualization

**Motivation:** Simulation studies that are carefully designed under realistic sample survey or census conditions can be used to evaluate the quality of new statistical methodology for Census Bureau data. Furthermore, new computationally intensive statistical methodology is often beneficial because it can require less strict assumptions, offer more flexibility in sampling or modeling, accommodate complex features in the data, enable valid inference where other methods might fail, etc. Statistical methods. Modeling also enables one to efficiently use all available information when producing estimates. Such studies can benefit from software for data processing, especially large data sets from nontraditional sources. Data visualizations can help reveal insights. Statistical disclosure avoidance methods are also developed and properties studied.

#### **Research Problems:**

- Systematically develop an environment for simulating complex sample surveys that can be used as a test-bed for new data analysis methods.
- Develop new methods for statistical disclosure control that simultaneously protect confidential data from disclosure while enabling valid inferences to be drawn on relevant population parameters.
- Develop models for the analysis of measurement errors in Demographic sample surveys (e.g., Current Population Survey or the Survey of Income and Program Participation).
- Investigate noise infusion and synthetic data for statistical disclosure control.

#### **Current Subprojects:**

- Development and Evaluation of Methodology for Statistical Disclosure Control (Nayak)
- The Ranking Project: Methodology Development and Evaluation (Wright, Klein/FDA, Wieczorek/Colby College, Yau)

#### **Potential Applications:**

- Simulating data collection operations using Monte Carlo techniques can help the Census Bureau make more efficient changes.
- Use noise multiplication or synthetic data as an alternative to top coding for statistical disclosure control in publicly released data. Both noise multiplication and synthetic data have the potential to preserve more information in the released data over top coding.
- Rigorous statistical disclosure control methods allow for the release of new microdata products.
- Using an environment for simulating complex sample surveys, statistical properties of new methods for missing data imputation, model-based estimation, small area estimation, etc. can be evaluated.
- Model-based estimation procedures enable efficient use of auxiliary information (for example, Economic Census information in business surveys), and can be applied in situations where variables are highly skewed and sample sizes are not sufficiently large to justify normal approximations. These methods may also be applicable to analyze data arising from a mechanism other than random sampling.
- Variance estimates and confidence intervals in complex sample surveys can be obtained via the bootstrap.
- Modeling approaches with administrative records can help enhance the information obtained from various sample surveys.

#### Accomplishments (October 2018-September 2020):

- Developed and published visualizations for comparing populations.
- Developed and published theory and a visualization for expressing uncertainty and an overall ranking of populations.
- Conducted research and published results on randomized response methods for protecting respondent's privacy and data confidentiality.

#### Short-Term Activities (FY 2021):

- Continue development of new methodology for statistical disclosure control and to evaluate properties of new and existing methods.
- Improve visualizations for comparing populations and for overall rankings of populations.

#### Longer-Term Activities (beyond FY 2021):

• Study ways of quantifying the privacy protection/data utility tradeoff in statistical disclosure control.

- Create an environment for simulating complex aspects of economic/demographic sample surveys.
- Develop methodology for quantifying uncertainty in statistical rankings, and refine visualizations.

#### Selected Publications:

- Klein, M., Wright, T., and Wieczorek, J. (2020). "A Joint Confidence Region for an Overall Ranking of Populations," *Journal of the Royal Statistical Society, Series C, 69, Part 3*, 589-606.
- Klein, M.D., Zylstra, J., and Sinha, B.K. (2019). "Finite Sample Inference for Multiply Imputed Synthetic Data under a Multiple Linear Regression Model," *Calcutta Statistical Association Bulletin*. https://doi.org/10.1177/0008068318803814
- Wright, T., Klein, M., and Wieczorek, J. (2019). "A Primer on Visualizations for Comparing Populations Including the Issue of Overlapping Confidence Intervals," *The American Statistician*, Vol 73, No. 2, 165-178.
- Chai, J. and Nayak, T.K. (2018). "A Criterion for Privacy Protection in Data Collection and Its Attainment via Randomized Response Procedures," *Electronic Journal of Statistics*, 12, 4264-4287.
- Klein, M. and Datta, G. (2018). "Statistical Disclosure Control via Sufficiency under the Multiple Linear Regression Model," *Journal of Statistical Theory and Practice*, 12(1), 100-110.
- Nayak, T.K., Zhang, C., and You, J. (2018). "Measuring Identification Risk in Microdata Release and Its Control by Postrandomisation," *International Statistical Review*, 86(2), 300-321.
- Moura, R., Klein, M., Coelho, C., and Sinha, B. (2017). "Inference for Multivariate Regression Model Based on Synthetic Data Generated under Fixed-Posterior Predictive Sampling: Comparison with Plug-in Sampling," *REVSTAT – Statistical Journal*, 15(2): 155-186.
- Klein, M. and Sinha, B. (2016). "Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data under the Multivariate Normal and Multiple Linear Regression Models," *Journal of Privacy and Confidentiality*, 7: 43-98.
- Klein, M. and Sinha, B. (2015). "Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models," *Sankhya B: The Indian Journal of Statistics*, 77-B, 293-311.
- Klein, M. and Sinha, B. (2015). "Likelihood-Based Inference for Singly and Multiply Imputed Synthetic Data under a Normal Model," *Statistics and Probability Letters*, 105, 168-175.
- Klein, M. and Sinha, B. (2015). "Likelihood-Based Finite Sample Inference for Synthetic Data Based on Exponential Model," *Thailand Statistician: Journal of The Thai Statistical Association*, 13, 33-47.
- Wright, T., Klein, M., and Wieczorek, J. (2014). "Ranking Populations Based on Sample Survey Data," *Center for Statistical Research and Methodology Directorate Research Report Series (Statistics #2014-12)*. U.S. Census Bureau. Available online: http://www.census.gov/srd/papers/pdf/rrs2014-12.pdf.
- Klein, M., Lineback, J.F., and Schafer, J. (2014). "Evaluating Imputation Techniques in the Monthly Wholesale Trade Survey," *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association.
- Klein, M., Mathew, T., and Sinha, B. (2014). "Noise Multiplication for Statistical Disclosure Control of Extreme Values in Lognormal Regression Samples." *Journal of Privacy and Confidentiality*, 6, 77-125.
- Klein, M., Mathew, T., and Sinha, B. (2014). "Likelihood Based Inference under Noise Multiplication," *Thailand Statistician: Journal of The Thai Statistical Association*, 12, 1-23.
- Wright, T., Klein, M., and Wieczorek, J. (2013). "An Overview of Some Concepts for Potential Use in Ranking Populations Based on Sample Survey Data," *The 59th International Statistical Institute World Statistics Congress*, Hong Kong, China.
- Klein, M. and Sinha, B. (2013). "Statistical Analysis of Noise Multiplied Data Using Multiple Imputation," *Journal of Official Statistics, 29*, 425-465.
- Klein, M. and Linton, P. (2013). "On a Comparison of Tests of Homogeneity of Binomial Proportions," *Journal of Statistical Theory and Applications*, 12, 208-224.
- Klein, M., Mathew, T., and Sinha, B. (2013). "A Comparison of Statistical Disclosure Control Methods: Multiple Imputation versus Noise Multiplication." *Center for Statistical Research and Methodology, Research and Methodology Directorate Research Report Series (Statistics #2013-02)*. U.S. Census Bureau. Available online: http://www.census.gov/srd/papers/pdf/rrs2013-02.pdf.
- Shao, J., Klein, M., and Xu, J. (2012). "Imputation for Nonmonotone Nonresponse in the Survey of Industrial Research and Development," Survey Methodology, 38, 143-155.
- Klein, M. and Wright, T. (2011). "Ranking Procedures for Several Normal Populations: An Empirical Investigation," *International Journal of Statistical Sciences*, 11, 37-58.
- Nayak, T.K., Sinha, B., and Zayatz, L. (2011). "Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection," *Journal of Official Statistics*, 27 (3), 527-544.
- Sinha, B., Nayak, T.K., and Zayatz, L. (2011). "Privacy Protection and Quantile Estimation from Noise Multiplied Data," *Sankhya*, Ser. B, 73, 297-315.
- Klein, M. and Creecy, R. (2010). "Steps toward Creating a Fully Synthetic Decennial Census Microdata File," *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association.

Contact: Tommy Wright, Tapan Nayak, Bimal Sinha, Nathan Yau

**Funding Sources for FY 2021:** 0331 – Working Capital Fund / General Research Project Various Decennial, Demographic, and Economic Projects

#### Cross-Cutting Statistical General Research Priorities (Study/Working Group Members)

A. Design and Analysis of Sample Surveys around Administrative or Commercial Observational Databases, or based on Web (Opt-in) Data-Collections: specific analysis and focus could be directed at the Census Bureau Contact Frame (MAF subset with telephone and email addresses) as used e.g. in the current Household Pulse Survey.

#### LEADER: Eric Slud

(FIRST Choice: Darcy, Chad, Eric, Michael I., Mary, Soumen, Dan, Tommy, Tapan; SECOND Choice: Anindya)

This initiative involves new methodology including modeling, to support Census Bureau efforts to design sample surveys and censuses in the future around special national lists as frames. Such lists may be convenient because of auxiliary data they contain, such as administrative records, or because they refer to address lists with auxiliary validated contact information such as telephone numbers or email/IP addresses. The CSRM effort includes descriptive statistical summaries of the predictive characteristics of membership on one or more lists of these types, leading to the development of effective predictive models to be used in future designs in tandem with the general MAF frame.

#### B. Optimization-based (single-stage) approaches to Weight-adjustment for Probability and Nonprobability Samples.

#### LEADER: Emanuel Ben-David (FIRST Choice: Emanuel; SECOND Choice: Mary, Isaac; OTHER Choice: Eric, Bimal)

Several of the Census Bureau's most important household surveys produce survey weights after several (up to 15!) successive difficult-to-document stages of adjustment or poststratification, with the result that the adjustments made in early stages are somewhat distorted in later stages. Methodology exists to do such poststratification in the form of Generalized Raking or Calibration by an optimization approach to minimize the degree of adjustment of base weights while ensuring exact or approximate conformity with calibration constraints to adhere to Population Estimates or other external-source totals for key variables. Considering the important application of Census Bureau surveys by survey methodologists to calibrate their own surveys for other purposes, this optimization-based approach would at the same time be easier to document and would maintain better simultaneous conformity with population controls than current methods.

## C. Research on model-based imputation incorporating (nonrandom) Hot-deck values as covariates, leveraging descriptive analyses of differences between the hot-deck donor universe and general population.

#### LEADER:

#### (FIRST Choice: Bimal; SECOND Choice: Darcy, Eric, Soumen, Anindya, Jun; OTHER Choice: Kim)

The Census Bureau relies throughout its household surveys on whole-unit and single-item imputation methodology based on hot-deck algorithms to impute or allocate missing data from data supplied nearby responding units. Attempts to update these methods with model-based improvements have generally failed, at least in part because demographic predictive variables omit important neighborhood information that is obtained from nearby donor units. Taking such donor information into account within predictive models, instead of using it directly in imputation, is an approach that has not been adequately explored, and that the Census Bureau is uniquely situated to implement properly. Research along these lines would clarify the differences between single and joint distributions of donor versus general-population household variables, and could improve the assessment and representativeness of joint distributions of variables in microdata, which are always partially imputed.

D. Development of Model Diagnostics and Cross-validation methods for Imputation and Small Area Estimation models.

#### LEADER: Carolina Franco

## (FIRST Choice: Ryan, Gauri, Isaac, Carolina, Kyle; SECOND Choice: Jerry, Kim;; OTHER Choice: Maria, Soumen, Darcy)

Throughout Census Bureau research efforts, model-based methods for response propensity prediction, for unit and item imputation, and for small-area estimates are impeded by the lack of systematic methodology for assessment involving ground truth. Ad hoc model diagnostics generally reveal only the differences between the results from competing models. Methods of cross-validation – currently under-developed in small-area and survey-sampling literature, would improve the Census Bureau's ability to ensure quality of released data, supported by increased use of post-enumeration survey results

from the decennial census.

#### E. Development of Survey and Sampling Microsimulation utility for applications to Nondisclosure (Synthetic Data), and to Testbeds for Model- and Design-based methods in Variance Estimation and (area- and unit-level) Small Area Estimation.

#### LEADER: Jerry Maples

(FIRST Choice: Jerry; SECOND Choice: Isaac, Carolina, Osbert, Joe, Tapan: OTHER Choice: Soumen, Gauri)

A system for microsimulation of artificial-population survey and census data, in the context both of household and economic surveys, would have at least two important ongoing applications: (i) as test-beds for current and new modelbased methods for producing imputations and special-purpose and small area data, and (ii) in the development of new methods for the release of partially synthetic data products whose nondisclosure properties and variability can be documented scientifically. Work along this line is already underway for some Economic surveys and for SAIPE testing.

#### F. Joint Time-Series/Spatial and Sampling Estimation Models and Diagnostics.

#### LEADER:

(FIRST Choice: Anindya, Kim, Osbert, Patrick, Jim; SECOND Choice: Ryan, Gauri, Joe, Dan; OTHER Choice: Carolina, Soumen)

Time series expertise could be leveraged toward providing specialized data tabulations, customized to special-purpose time periods and small areas, if research were expanded on models and estimation methods jointly incorporating time series and sampling errors. There is considerable expertise in CSRM on time series forecasting, on demographic and on small area modeling and benchmarking. Development of time series methods for custom tabulations would promise new custom data products as well as new tools for assessing small-area estimates produced throughout the Census Bureau.

G. Other items (?) e.g., Methods of Assessing Variability of Census or Survey Totals based on Noise-Infused Data; random-based causal inference for complex experiments; design and analysis of experiments on non-response using sequential regression models; entity resolution, visualizations; small area estimation/longitudinal studies. (FIRST Choice: Thomas, Andrew, Beka, Jim, Nathan; SECOND Choice: Emanuel, Kyle, Chad, Michael I., Bimal, Tommy)

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-3 Filed 04/20/21 Page 1 of 27

## Exhibit 3

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-3 Filed 04/20/21 Page 2 of 27

# Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the 2018 End-to-End Test (Continued)

Simson L. Garfinkel Chief, Center for Disclosure Avoidance Research U.S. Census Bureau

2017 Census Scientific Advisory Committee Fall Meeting Suitland, MD 11:00AM September 15, 2017



## Acknowledgments

This presentation incorporates work by:

- Dan Kifer (Scientific Lead)
- John Abowd (Chief Scientist)
- Tammy Adams, Robert Ashmead, Aref Dajani, Jason Devine, Michael Hay, Cynthia Hollingsworth, Meriton Ibrahimi, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Ned Porter, Anne Ross and William Sexton



## Outline

Motivation

Differentially private 2020 Disclosure Avoidance System

- High-level goals
- Flow diagrams
- Query examples
- Conclusion



# **Motivation:**

To protect the privacy of individual survey responses

2010 Census:

- 5.6 billion independent tabular summaries published.
- Based on 308 million person records

Database reconstruction (Dinur and Nissim 2003) is a serious disclosure threat that all statistical tabulation systems from confidential data must acknowledge.

The confidentiality edits applied to the 2010 Census were not designed to defend against this kind of attack.



# The Disclosure Avoidance Subsystem (DAS) implements the privacy protections for the decennial Census.

## Features of the DAS:

- Operates on the edited Census records
- Designed to make Census records safe to tabulate




# The 2000 and 2010 Disclosure Avoidance Systems relied on swapping households:

#### Advantages of swapping:

- Easy to understand
- Does not affect state counts if swaps are within a state
- Can be run state-by-state
- Operation is "invisible" to rest of Census processing

Disadvantages:

- Does not provide formal privacy guarantees
- Does not protect against database reconstruction attacks
- Privacy guarantee relies on lack of external data



State "X"

# The 2000 and 2010 Disclosure Avoidance System operated as a filter, on the Census Edited File:





# The 2020 Census disclosure avoidance system will use differential privacy to defend against a reconstruction attack,

## Differential privacy provides:

- Provable bounds on the accuracy of the best possible database reconstruction given the released tabulations.
- Algorithms that allow policy makers to decide the trade-off between accuracy and privacy.





# The 2020 Disclosure Avoidance System also operates as a filter... but it's much more visible.



# The 2020 Disclosure Avoidance System relies on infusing formally private noise.

Advantages of noise infusion with formal privacy:

- Easy to understand
- Provable and *tunable* privacy guarantees
- Privacy guarantees do not depend on external data
- Protects against database reconstruction attacks
- Privacy operations are *composable*

### Disadvantages:

- Entire country must be processed at once for best accuracy
- Every use of private data must be tallied in the privacy loss budget





### Differentially Private Disclosure Avoidance System: Requirements

DAS must be able to read the Census Edited File (CEF):

- CEF must be exactly specified and contain all information necessary for all tabulation recodes
- CEF must be kept confidential after DAS runs (as it was for historical censuses)
   DAS must generate the Microdata Detail File (MDF):
- Must contain all information that appears in any publicly released table (e.g. PL94-171, SF1, SF2)
- Should not contain *any information* that does *not* appear in a publicly released table
- May be publicly released (in whole or in part)

Non-functional requirements:

- The disclosure avoidance system must provably move information from the CEF to PL94/SF1/SF2 with an adjustable total privacy-loss budget
- The source code and parameters for the DAS will be made publicly available



### Why generate a differentially private MDF?

- Familiar to internal and external stakeholders
- Operates with tabulation system to produce PL-94 and SF-1 tabulations
- Guarantees population totals (voting age, non-voting age, householder) exact at all levels of geography
- Consistency among query answers



# Some queries must be privacy preserving. Some queries must be exact ("invariant")

# Specific PL-94 queries must be exact:

- Block population
- Block voting age population
- Block householders & vacancies

—per 2000 Department of Justice letter to the Director, Kenneth Prewitt

Other PL-94 and SF-1 queries will not be exact:

- Age distribution under 18
- Age distribution 18 and over
- Race and ethnicity distribution
- Household relationship distribution
- Household ownership distribution

Final privacy-loss budget determined by Data Stewardship Executive Policy Committee (DSEP) with recommendation from Disclosure Review Board (DRB)



## How the 2020 System Works: High-level Overview

The new system is similar to swapping, with key differences:

- Every record in the population may be modified But modifications are bounded by DOJ policy and global privacy budget.
- Records in the tabulation data have no exact counterpart in the confidential data

There is no one-to-one mapping between CEF and MDF records. But there are the same number of records for every block.

 Explicitly protected tabulations (PL-94 and SF-1) have provable, public accuracy levels

2020 will publish the algorithms, the parameters and the accuracy of the tabulations.



## **Proposed "Top-Down" Algorithm**

Spend ε<sub>1</sub> privacy-loss budget

### 2 x 255 x 17 x 115

National table of

**US** population

Sex: Male / Female Race + Hispanic: 255 possible values Relationship to Householder: 17

Age: 0-114



National table with all 500,000 cells

filled, structural zeros imposed with

accuracy allowed by  $\boldsymbol{\varepsilon}_1$ 

2 x 255 x 17 x 115

Reconstruct individual micro-data without geography

325,000,000 records





U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov

## **State-level**



State-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend  $\epsilon_2$ privacy-loss budget Target state-level tables required for best accuracy for PL-94 and SF-1 Exact state voting-age, non-voting age, and householder counts as enumerated.





U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov



Construct best-fitting individual micro-data with state geography

325,000,000 records now including state identifiers



## **County-level**



Target county-level tables required for best

accuracy for PL-94 and SF-1

Exact county voting-age, non-voting age,

and householder counts as enumerated.

County-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend **ε**<sub>3</sub> privacy-loss budget



U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov Construct best-fitting individual micro-data with state and county geography

325,000,000 records now including state and county identifiers



## **Census tract-level**



Tract-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend **ε**<sub>4</sub> privacy-loss budget Target tract-level tables required for best accuracy for PL-94 and SF-1

Exact tract voting-age, non-voting age, and householder counts as enumerated.





U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov



Construct best-fitting individual micro-data with state, county, and tract geography

325,000,000 records now including state, county, and tract identifiers



## **Block-level**



Block-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1 Spend **ε**<sub>5</sub> privacy-loss budget Block tract-level tables required for best accuracy for PL-94 and SF-1 Exact block voting-age, non-voting age, and householder counts as enumerated.





U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov



Construct best-fitting individual micro-data with state, county, tract and block geography

325,000,000 records now including state, county, tract identifiers



# **MDF for tabulating**

Construct best-fitting individual micro-data with state, county, tract and block geography

325,000,000 records now including state, county, tract, and block identifiers



MDF used for tabulating PL-94, SF-1



# **MDF for tabulating**

#### How accurate is the MDF?



**Disclosure Avoidance Certificate** 

- Certifies that the DAS passed tests
- Reports the accuracy of the MDF
- Requires  $\boldsymbol{\epsilon}_A$



Construct best-fitting individual micro-data with state, county, tract and block geography

325,000,000 records now including state, county, tract, and block identifiers



MDF used for tabulating PL-94, SF-1

### **Operational Decisions**

Set total privacy loss budget: **ɛ** 

• Ensure that  $\boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2 + \boldsymbol{\varepsilon}_3 + \boldsymbol{\varepsilon}_4 + \boldsymbol{\varepsilon}_5 + \boldsymbol{\varepsilon}_A = \boldsymbol{\varepsilon}$ 

Privacy Loss Budget	
No Privacy	No Accuracy
Accuracy Allocation	
PL94	SF1

Within each stage, allocate privacy-loss budget between:

- PL-94
- Parts of SF-1 not in PL-94

These are policy levers provided by the system.

Levers are set by the Data Stewardship Executive Policy Committee



#### Inputs Used by the Development Team

Lists of matrices in technical documentation express core queries in the workload

- PL94: https://www.census.gov/prod/cen2010/doc/pl94-171.pdf
- SF1: https://www.census.gov/prod/cen2010/doc/sf1.pdf
- SF2: https://www.census.gov/prod/cen2010/doc/sf2.pdf

## Over 1,000 pages of edit specifications for 2010 CEF Uncurated tabulation recode programs



# We are creating

# A *framework* for Disclosure Avoidance Systems:

- Development & Test Mode
- Production Mode

#### Testing Systems:

- DAS0 100% accuracy, no privacy (No disclosure avoidance)
- DAS1 100% privacy, no accuracy
- DAS2 "bottom-up" engine

#### **Operational System:**

DAS3 — "top-down" engine





#### Plans for the 2018 End-to-End Test

The 2018 End-to-End test will incorporate differential privacy

Likely DAS2 — Bottom-up algorithm

Only the prototype PL94-171 files will be produced

No decisions yet regarding the privacy-loss budget or accuracy level Questions?



#### Reference

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.



Case 3:21-cv-00211-RAH-ECM-KCN Document 94-4 Filed 04/20/21 Page 1 of 4

#### Exhibit 4

 To:
 Kathleen M Styles (CENSUS/ADDC FED)[kathleen.m styles@census.gov]...

 From:
 James Whitehome (CENSUS/ADDC FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP

 (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=010B27ED5B5944CDA73EB75917D5A0DD-WHITEHORNE,]

 Sent:
 Tue 7/7/2020 1:19:57 PM (UTC)

 Subject:
 Fw: Short video to help with Wednesday's discussion

FYI

#### \*\*\*\*

James Whitehorne, Chief Redistricting & Voting Rights Data Office/ADDC/HQ U.S. Census Bureau O: 301-763-4039 | M: 202-263-9144 census.gov | census.gov/rdo | @uscensusbureau Shape your future. START HERE > 2020census.gov

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov> Sent: Tuesday, July 7, 2020 9:15 AM To: Ron S Jarmin (CENSUS/DEPDIR FED) <Ron.S.Jarmin@census.gov>; Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>; James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>; Christa D Jones (CENSUS/DEPDIR FED) <Christa.D.Jones@census.gov>; Ali Mohammad Ahmad (CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>; Enrique Lamas (CENSUS/DEPDIR FED) <Enrique.Lamas@census.gov>; Albert E Fontenot (CENSUS/ADDC FED) <Albert.E.Fontenot@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov> Subject: Re: Short video to help with Wednesday's discussion

Michael should say clearly at the start of his presentation that the fifty state resident population totals (plus DC and PR, but they are irrelevant) are outside the differential privacy system. They will be released exactly as enumerated. That decision has already been memorialized in a 2020 Decision memo that Michael should find and you should send ahead to the Secretary. If asked, I can go into the reasoning, but I doubt that it will come up again if we make it clear from the outset.

The really difficult thing to explain is why the population invariant cannot be extended below the state level without breaking both the privacy protection (of location) and the TopDown algorithm itself. It breaks the privacy protection because it means that no formal guarantees (the things that give differential privacy its confidentiality protection properties) exist for the protection of the household's address, a key PII component that we must protect. The absence of noise in the 2010 block population counts is the single biggest contributor to our (lower bound) estimate of 52 million reidentified households. Even our swapping algorithms would have had to add noise to those counts in 2020. More population count invariants break TopDown because the proof of existence of a solution to the post-processing algorithm, in particular the rounding step to produce microdata, relies on an integer programming property called total unimodularity (TUM). We have struggled to incorporate the housing unit and group quarters invariants (which are not population count invariants) within the TUM constraaints. I intend to tell DSEP in September that if any population invariants are extended below the state level, I cannot guarantee that the 2020 DAS will work because that will break the current failsafe that ensures that we get a solution. In contrast, DSEP can set the global privacy-loss budget at any level that it is comfortable with and assign as much of it to population counts as it wishes. Even with a PLB of 4 (as in the demonstration data products) an allocation of 30% to population counts produces a mean absolute error in populations for counties of 16. Allocating 75% of that PLB to population drives this error to less than 9. (look here: https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosureavoidance-system/2020-06-01-metrics-update.pdf?#) We can drive it down as much as we want, but each time we do, the privacy protection of location is further compromised.

Thanks,

John M. Abowd, PhD, Associate Director and Chief Scientist Research and Methodology U.S. Census Bureau O: <u>301-763-5880</u> M: simulring on cell From: Ron S Jarmin (CENSUS/DEPDIR FED) <Ron.S.Jarmin@census.gov>
Sent: Tuesday, July 7, 2020 8:49 AM
To: Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>; James Whitehorne (CENSUS/ADDC FED)
<James.Whitehorne@census.gov>; Christa D Jones (CENSUS/DEPDIR FED) <Christa.D.Jones@census.gov>; Ali Mohammad Ahmad
(CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>; Enrique Lamas (CENSUS/DEPDIR FED) <Enrique.Lamas@census.gov>; Albert E
Fontenot (CENSUS/ADDC FED) <Albert.E.Fontenot@census.gov>; John Maron Abowd (CENSUS/ADRM FED)
<john.maron.abowd@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>
Subject: Re: Short video to help with Wednesday's discussion

Per his question, we need to stress decisions on key parameters not yet made. Also, a summary of alternative approaches and why we went in this direction. Why we settled on top down vs others.

Note I will need to jump off the call at 10 to get on another.

On Jul 6, 2020, at 6:19 PM, Ron S Jarmin (CENSUS/DEPDIR FED) <Ron.S.Jarmin@census.gov> wrote:

FYI

Begin forwarded message:

From: Wilbur Ross <wlr@doc.gov> Date: July 6, 2020 at 6:13:54 PM EDT To: "Ron S Jarmin (CENSUS/DEPDIR FED)" <Ron.S.Jarmin@census.gov> Cc: "Kelley, Karen (Federal)" <KKelley@doc.gov>, "Steven Dillingham (CENSUS/DEPDIR FED)" <steven.dillingham@census.gov> Subject: Re: Short video to help with Wednesday's discussion

ThNk you. I have no doubt that we want to use differential privacy, but as I understand it there is a trade off between more privacy and potentially less accuracy. I also understNd that as a result you experimented with T least a few different approaches. Since apportionment is a hot topic, I would like to understand how the different models might affect it and how you came to the conclusion that you did. WLR Sent from my iPhone

On Jul 6, 2020, at 4:41 PM, Ron S Jarmin (CENSUS/DEPDIR FED) <<u>Ron.S.Jarmin@census.gov</u>> wrote:

Good afternoon Secretary Ross,

Here's a video from "Minute Physics" on differential privacy that we collaborated on a while back.

https://www.youtube.com/watch?v=pT19VwBAqKA&pbjreload=101

This might help as background for Wednesday's discussion. While differential

DOC AL 0215160

Case 3.21 - CV-90291 - RAFE - Engla Consus for 2020, it's not completely 26/21 o bage 4 of 4 Census Bureau. We employed it for privacy protection on some products I believed you've been briefed on including:

OnTheMap for Emergency Management

 (https://onthemap.ces.census.gov/em/)

 Post Secondary Employment Outcomes (PSEO)

 (https://lehd.ces.census.gov/data/pseo\_explorer.html?type=earnings&compa re=postgrad&state=08&institution=001370&degreelevel=05&gradcohort=000
 0-3&filter=50&program=5202,4201)

 Veterans' Employment Outcomes

 (https://lehd.ces.census.gov/applications/veo/occupation/detail/)

Without differential privacy, these products would not be feasible as they would violate confidentiality or be so coarsened as to be useless.

Let me know if you have questions.

Thanks

Ron S Jarmin, PhD., Deputy Director

U.S. Census Bureau o: 301-763-1858 | m: 301-980-8140 census.gov | @uscensusbureau Shape your future. START HERE > <u>2020census.gov</u> Case 3:21-cv-00211-RAH-ECM-KCN Document 94-5 Filed 04/20/21 Page 1 of 3

#### Exhibit 5

To: John Maron Abowd (CENSUS/ADRM FED)[john maron.abowd@census.gov] Cc: Robert Stenkiewicz (CENSUS/CED FED)[joben.stenkiewicz@census.gov]; Simson L Garinkel (CENSUS/ADRM FED)[simson.l.garfinkel@census.gov]; Michael B Hawes (CENSUS/CED FED)[michael.b.hawes@census.gov] From: Gary Linus Benedetto (CENSUS/CED FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=9E0D0BFF599448BD94D5FE38D6850E4E-BENEDETTO,] Sent: Thur 7/30/2020 8:01:02 PM (UTC) Subject: Re: Please fix this 20200728 Disclosure Modernization Update-JMA-GB.docx

In that case, I have attached a version that has incorporated Rolando's feedback into the ACS section, Ian's feedback into the ECON section, and a few edits from me (including the one Michael suggested in our meeting with Rob about this last week). I am now passing it to Michael to update the 2020 Decennial section, as well as another once over.

Thanks, Gary

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov> Sent: Tuesday, July 28, 2020 1:01 PM To: Gary Linus Benedetto (CENSUS/CED FED) <Gary.Linus.Benedetto@census.gov> Cc: Robert Sienkiewicz (CENSUS/CED FED) <robert.sienkiewicz@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov> Subject: Re: Please fix this

Michael is the one who really needs it. He is fact checking a HDSR article that quotes heavily from the older paper, including the garbage about being legally required to hold certain populations invariant.

Thanks,

John M. Abowd, PhD, Associate Director and Chief Scientist Research and Methodology U.S. Census Bureau O: <u>301-763-5880</u> M: simulring on cell <u>census.gov</u> | <u>@uscensusbureau</u> Shape your future. START HERE > <u>2020census.gov</u>

From: Gary Linus Benedetto (CENSUS/CED FED) <Gary.Linus.Benedetto@census.gov>
Sent: Tuesday, July 28, 2020 12:51 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Cc: Robert Sienkiewicz (CENSUS/CED FED) <robert.sienkiewicz@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED)
<simson.l.garfinkel@census.gov>; Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>
Subject: Re: Please fix this

I have sent along the section requests to Michael, Rolando, and Ian. I have some vacation planned for 8/5-7, so I will try to have something passed along before then, but it is possible it might come a little after 8/7.

Best, Gary From: John Maron Abowd (CENSUS/ADRM FED) <iohn maron abowd@census.gov> Filed 04/20/21 Page 3 of 3 Sent: Tuesday, July 28, 2020 11:32 AM To: Gary Linus Benedetto (CENSUS/CED FED) <Gary.Linus.Benedetto@census.gov> Cc: Robert Sienkiewicz (CENSUS/CED FED) <robert.sienkiewicz@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov> Subject: Please fix this

I want this paper so that it replaces the UNESCO piece, not "updates" with a lot of the incorrect reasoning in the original preserved. Aref started but you cannot assign this to him for further input. Assign it to Michael, Roberto and Ian to update for the 2020 Census, ACS and Econ Census, resp. Remove authors who made no contribution. You, Simson and I are the lead authors, remainder in alphabetical order. I rewrote large chunks of the original, but much of what is said here is now out of date. After your review pass to Simson; after Simson, pass to Michael, then send to me.

Due by Friday, August 7th.

Thanks,

John M. Abowd, PhD, Associate Director and Chief Scientist Research and Methodology U.S. Census Bureau O: <u>301-763-5880</u> M: simulring on cell <u>census.gov</u> | <u>@uscensusbureau</u> Shape your future. START HERE > <u>2020census.gov</u> Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 1 of 61

#### Exhibit 6

#### 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; GINA RAIMONDO, *et al.*,

Defendants.

CASE NO. 3:21-cv-00211-RAH-ECM-KCN

#### **DECLARATION OF DR. STEVEN RUGGLES**

I, DR. STEVEN RUGGLES, 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 over the age of 18 and competent to make this declaration.

2. I am the Regents Professor of History and Population Studies at the University of Minnesota and I direct the University's Institute for Social Research and Data Innovation.

3. I received a PhD from the University of Pennsylvania in 1984 with a focus on historical demography, and subsequently undertook postdoctoral training in demography at the University of Wisconsin-Madison.

4. I have over 35 years of experience in the areas of demography and populations studies, including the study and use of both U.S. and international census data.

5. I have been the principal investigator on various census related research projects including, but not limited to, a study on the Implications of Differential Privacy on Decennial Census Data Access and Utility, which received funding from the Sloan Foundation.

6. My research focus is on methods of census curation, census data dissemination, the history of the U.S. Census, disclosure control in statistical databases (including differential privacy), and long-run demographic trends. I have published extensively on these topics.

7. I am currently the Director of the Institute for Social Research and Data Innovation and the Director of the International Integrated Public Use Microdata Series ("IPUMS") Center for Data Integration. Between 2000 and 2016, I was the Director of the Minnesota Population Center.

8. I have never previously been qualified as an expert witness in federal court.

9. I have contributed to several *Amici Curiae* briefs related to recent Census Bureau litigation.

2

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 4 of 61

10. Plaintiffs requested that I assess the Census Bureau's decision to use of differential privacy, including assertions made by Dr. Abowd in his declaration. To that end, I analyzed the following: (1) differential privacy's use as a disclosure avoidance technique; (2) the threat, or lack thereof, posed by database reconstruction attacks; (3) the fitness for use of census data after differential privacy is applied; and (4) differential privacy's use in the context of historic Census Bureau practice.

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

12. 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 A. My curriculum vitae is attached to the expert report as Appendix 1.

13. My curriculum vitae lists, among other things, my qualifications, and a list of all publications published over at least the last ten years. I have neither testified in federal court nor have I given a deposition.

14. 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: April 20, 2021

Sten Ryle

Dr. Steven Ruggles

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 5 of 61

### Appendix A

#### 1. Introduction and Qualifications

The plaintiffs in Alabama vs the Department of Commerce have retained my services to analyze various assertions that have been made by the defendant in this case. Specifically, I have reviewed the pleadings, including but not limited to, the complaint, defendants' answers to plaintiffs' requests for admissions, plaintiffs' motion for preliminary injunction, and defendants' response. The opinions I set forth below are based on these pleadings and informed by my work in the field of demography as explained below and in my curriculum vita. I am being compensated 300 dollars per hour for this work.

I am Regents Professor of History and Population Studies at the University of Minnesota and I direct the University's Institute for Social Research and Data Innovation. I received a PhD from the University of Pennsylvania in 1984 with a focus on historical demography, and subsequently undertook postdoctoral training in demography at the University of Wisconsin-Madison.

I developed the world's largest census database, known the Integrated Public Use Microdata Series (IPUMS), which has been used by over 200,000 investigators to conduct demographic and economic research. I have authored more than 100 publications on methods of census curation, census data dissemination, the history of the U.S. Census, disclosure control in statistical databases (including differential privacy), and long-run demographic trends. My research has appeared in, among other outlets, *Demography, Population and Development Review, Population Studies, American Sociological Review, Annual Review of Sociology, American Historical Review, Journal of American History,* and *Privacy in Statistical Databases*. I have served as Principal Investigator or Co-Principal Investigator on 57 grants to digitize, curate, analyze, and disseminate census data, including a Sloan Foundation grant to investigate the impact of differential privacy on the accuracy of the census.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 7 of 61

My work has been recognized by the Sharlin Award (Social Science History Association), the Goode Award (American Sociological Association), the Lapham Award (Population Association of America), and the Miller Award (Inter-university Consortium for Political and Social Research). I have been elected President of three national scholarly associations: The Population Association of America, the Association of Population Centers, and the Social Science History Association.

I have served on several relevant national committees, including the Census Scientific Advisory Committee (U.S. Census Bureau); the Advisory Committee on Cyberinfrastructure (National Science Foundation); the Committee on Education for Digital Curation (National Research Council Board on Research Data and Information); and the Advisory Committee for the Social, Behavioral, and Economic Sciences (National Science Foundation). I chaired the Working Group on Open Access to Data in the Social and Behavioral Sciences for the National Science Foundation.

I have participated in several *Amici Curiae* briefs related to recent Census Bureau litigation but have never previously served as an expert witness. These briefs are listed in my CV, attached hereto as Appendix 1.

#### 2. Overview of Census Bureau disclosure control

From 1970 through 2010, the Census Bureau used a variety of techniques, including table suppression (1970–1980), blank and impute (1990), and swapping (1990–2010) to protect the confidentiality of respondents. To implement these methods, the Bureau identified potentially disclosive variables and then found cells with small counts based on those variables. They then suppressed tables with these small counts or swapped households matched on key demographic characteristics between geographic units (McKenna 2018).

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 8 of 61

Traditional statistical disclosure control techniques introduced uncertainty into published data. Whole table suppression withheld information about certain aspects of the population. Swapping introduced error into some counts because households would not match on all demographic characteristics. The disclosure control methods used prior to 2020 did not, however, alter the counts of total population and voting age adults at any geographic level. Some noise was introduced on other characteristics, but the Census Bureau concluded that "the impact in terms of introducing error into the estimates was much smaller than errors from sampling, non-response, editing, and imputation" (McKenna 2018: 24).

The traditional Census Bureau disclosure control strategy has focused on ensuring that the responses of identified persons cannot be determined from census publications. The Census Bureau implemented 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 (Lauger, Wisniewski, and McKenna 2014). Indeed, *there is not a single documented case of anyone outside the Census Bureau revealing the responses of a particular identified person using data from the decennial census*.

The defendants deny that there is no documented case of outsiders identifying the responses of a particular identified person (see Defendants' Responses to Plaintiffs' First Request for Admissions, no. 6). They do not, however, document any such case of disclosure. In their sole justification for the denial, the defendants cite McKenna (2019), "U.S. Census Bureau Reidentification Studies." That citation is puzzling, since McKenna does not describe *any* reidentification attempts conducted outside the Census Bureau. Moreover, McKenna does not
#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 9 of 61

discuss any attempted reidentification of decennial census data. McKenna does describe an attempted attack on the American Community Survey, which concluded that just 0.005% of the population was vulnerable to identification. The great majority—78%—of the attempted identifications, however, were incorrect, and no identifications could be confirmed without access to internal identified data. McKenna's discussion therefore supports the statement that there is not a single documented case of anyone outside the Census Bureau uncovering the responses of a particular identified person using either the Decennial Census or the American Community Survey.

Despite the proven effectiveness of traditional statistical disclosure control, the Census Bureau has announced that it will be replaced for the 2020 census with an entirely new approach known as differential privacy. The use of differential privacy will reduce the accuracy of the census and may increase disclosure risk. Differential privacy is not designed to prevent the disclosure of the identities of census respondents and cannot guarantee absolute protection against such identification.

Implementations of differential privacy generally involves calculating cross-tabulations from "true" data and injecting noise drawn from a statistical distribution into the cells of the cross-tabulation. There are two significant consequences of this approach:

• The noise introduced into each cell is independent of the original value of the cell. Therefore, even if the noise is small relative to the average cell value, distortions in small cell values are often proportionally large. For example, the error introduced in the population of small towns can be proportionally large, sometimes exceeding 100% of the town's true population. • Simple random noise can produce logical inconsistencies, such as negative population counts or household counts that exceed population counts. If the data producer wishes to maintain logical consistency or preserve some noise-free counts, they must use a post-processing algorithm to adjust totals after noise injection, and this post-processing introduces additional types of error and systematic biases. In the preliminary Census Bureau demonstration datasets using differential privacy, such systematic biases are ubiquitous.<sup>1</sup>

Accordingly, differential privacy imposes high costs on the usability of data with no compensating benefits.

#### 3. Database reconstruction does not pose a realistic disclosure threat

The Census Bureau justifies the need for differential privacy by claiming that the confidentiality of census responses is threatened by "database reconstruction." Database reconstruction is a process for inferring individual-level responses from tabular data. John Abowd, the primary architect of the Census Bureau's new approach to disclosure control, argues that database reconstruction "is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared" (Abowd 2017).

<sup>&</sup>lt;sup>1</sup> To enable the research community to assess the consequences of differential privacy for the research and policy communities, the Census Bureau has released several demonstration datasets. The Census Bureau also released the source code that had been used to implement differential privacy, enabling investigators to experiment on their own. Over the past three years, multiple investigators seized these opportunities to understand the impact of differential privacy on census accuracy and usability. There have been several workshops and meetings devoted to the topic, including IPUMS Differential Privacy Workshop (August 15-16, 2019), the Harvard Data Science Review Symposium (October 25, 2019), the Committee on National Statistics Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations (December 11-12, 2019), and the 2020 Privacy in Statistical Databases conference (September 23-25, 2020). Additional work has appeared as working papers, as well as a few early publications. The following discussion draws on insights of this research.

Although Census Bureau staff members have repeatedly invoked database reconstruction to justify the use of differential privacy in public presentations, they have never, to my knowledge, produced a full description of their experiment, and some details remain obscure. There are no peer-reviewed publications explaining their methodology, and the experiment has not been replicated by outside experts.<sup>2</sup> Nevertheless, the various slide decks, blog posts, and tweets produced by the Census Bureau—now supplemented with Abowd Decl. App'x B—provide overwhelming evidence that the database reconstruction experiment failed to demonstrate a realistic disclosure risk. On the contrary, the database reconstruction exercise provides compelling evidence that even with a massive investment of time, resources, and computing power, it would be impossible for an outside attacker to infer the characteristics of a particular individual respondent from the published tabulations used for the 2010 census.

The Census Bureau conducted a database reconstruction experiment that attempted to determine the age, sex, race, and Hispanic or Non-Hispanic ethnicity for the population of each of the 6.3 million inhabited census blocks in the 2010 census. According to Abowd (2018a), the experiment confirmed that the individual-level census data "can be accurately reconstructed" using the published tabular census data. That assertion is false.

The "reconstructed" data produced by the experiment consists of rows of data identifying the age, sex, and race/ethnicity for each person in a hypothetical population of each census block; it does not include identifying information such as name, address, or Social Security number. Thus, for example, the hypothetical population of a given block could include a 26-year-old non-Hispanic white female. Contrary to the statements of the Census Bureau, the results demonstrated that they failed to reconstruct the individual-level characteristics of the population based on the

<sup>&</sup>lt;sup>2</sup> The experiment has been mentioned in several publications, e.g. Garfinkel, Abowd, and Powazek (2018) and Garfinkel, Abowd, and Martindale (2018), but these publications provide few details.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 12 of 61

published census tabulations. In fact, the Census Bureau found that *for most of their hypothetical population, there was not a single case in the real population that matched on block, age, sex, and race/ethnicity* (Abowd 2018b).

Among the minority of cases where the Census Bureau did find a match between their hypothetical population and a real person, most matches apparently occurred purely by chance.<sup>3</sup> Remarkably, the Census Bureau does not appear to have calculated how many matches would have been expected through chance alone. To investigate the matter, I constructed a simple simulation. I estimate that randomly chosen age-sex combinations would match someone on any given block 54.9% of the time, assuming the age, sex, and block size distributions from the 2010 census.<sup>4</sup> This means that, the Census Bureau would have been "correct" on age and sex 55% of the time even if they had never looked at the tabular data from 2010, and had instead just assigned ages and sexes to their hypothetical population at random.

This calculation does not factor in race or ethnicity, but because of high residential segregation most blocks are highly homogenous with respect to race and ethnicity. If we assign everyone on each block the most frequent race and ethnicity of the block, and then randomly

<sup>&</sup>lt;sup>3</sup> Abowd Decl. App'x B, Figure 1 shows that the "exact match rate" was positively associated with the number of people on the block: The larger the block, the more exact matches. Abowd reports that large blocks had three times the match rate of small blocks. Database reconstruction, however, ought to work best with small blocks, not large ones. The obvious explanation is that larger blocks have higher odds of including by chance any specific combination. of age, sex, race, and ethnicity. A block with 10 residents is unlikely by chance to include a 26-year old white non-Hispanic woman, but if the block has 1000 residents the odds are high. Abowd *Id* p. 3 attributes the strong positive association between block size and match rate to higher swap rates for small blocks. This explanation would, however, require an implausible level of swapping that is inconsistent with the low level of noise in the 2010 data (McKenna 2018).

<sup>&</sup>lt;sup>4</sup> To estimate the percentage of random age-sex combinations that would match someone on a block by chance, I generated 1000 simulated blocks and populated them with random draws from the 2010 single-year-of-age and sex distribution. The simulated blocks conformed to the population-weighted size distribution of blocks observed in the 2010 census. I then randomly drew 1000 new age-sex combinations and searched for each of them in each of the 1000 simulated blocks. In 54.9% of cases I found someone in the simulated block who exactly matched the random age-sex combination. The simulation source code and supporting data files are available at <a href="http://users.hist.umn.edu/~ruggles/censim.html">http://users.hist.umn.edu/~ruggles/censim.html</a>.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 13 of 61

choose age-sex combinations as described above, then 42.7% percent of cases in the hypothetical database would be expected to match on all four characteristics to a respondent on the same block. That is close to the 46.48% match rate for the reconstructed data reported in Abowd Decl. App'x B Table 1. Thus, despite the Census Bureau's enormous investment of resources and computing power, the 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.<sup>5</sup>

Undeterred by their failure to reconstruct most of the population correctly, the Census Bureau then assessed whether their hypothetical population shared characteristics with people who appeared in non-census sources. In particular, within each block they matched the age and sex of persons in the hypothetical population of each block to the age and sex of persons to "commercial data."<sup>6</sup> Thus, for example, they would search the commercial data for a real 26-year-old female to match the hypothetical person in the population reconstruction. A match on race or ethnicity was not required for this experiment. Once again, the results cast doubt on the reliability of the reconstructed data. In most cases, the hypothetical individuals constructed by the Census Bureau did not share the same age, sex, and block as *anyone* in the commercial data. In just 45% of cases was there at least one person in the commercial data who matched the age, sex and block number

<sup>&</sup>lt;sup>5</sup> Calculation of the race and ethnic predicted matches was based on U.S. Census Bureau (2011). The implications of the simulated data are reinforced by newly released data from the real census. Abowd Decl. App'x B ¶ 9 reports that in the real 2010 population, 44% of the population is unique within blocks with respect to age and sex, a figure entirely consistent with my finding that a randomly chosen age-sex combination would match someone on any given block 54.9% of the time. The simulated population was virtually identical to the real census population with respect to the frequency of unique respondents: I found that 45.2% of the simulated population was unique within the block with respect to age and sex.

<sup>&</sup>lt;sup>6</sup> The commercial data consists of marketing and financial data purchased from vendors including Experian, Targus, Veteran Service Group of Illinois, InfoGroup, and Melissa Data. The Census Bureau purchased the datasets to conduct the 2010 Census Match Study, which evaluated the utility of commercial records for census enumeration. See Rastogi and O'Hara (2012), U.S. Census Bureau (2021).

of at least one row of the hypothetical database. If the commercial data were accurate and complete, this is a *lower* match rate than one would expect purely by chance.

Among the cases where there was at least one person in the commercial database who matched the age, sex, and block of a row in the hypothetical population, the Census Bureau then harvested the names from the commercial database, and attempted to match them with names on the same block as enumerated in the 2010 census. Once again, the results highlighted the unreliability of the method: they found that just 38% of the names from the commercial database were actually present on the block.

Based on this exercise, the Census Bureau claimed to have successfully "re-identified" 16.85% (38% of 45%) of the population. That claim is irresponsible. Reidentification means confirming the identity of a particular individual and revealing their characteristics without reference to non-public internal census files. It would be impossible to positively identify the characteristics of any particular individual using the database reconstruction without access to nonpublic internal census information. Accordingly, the Census Bureau's database reconstruction experiment demonstrates that reidentification based on the published census tables is not feasible.

Abowd (2018b) acknowledged that the database reconstruction experiment demonstrates that "the risk of re-identification is small." Abowd has now retracted that statement (Abowd Decl. ¶83), but his supervisor has not. Acting Director of the Census Bureau Jarmin actually went farther than Abowd, writing "The accuracy of the data our researchers obtained from this study is limited, and confirmation of re-identified responses requires access to confidential internal Census Bureau information … an external attacker has no means of confirming them" (Jarmin 2019).

The database reconstruction exercise demonstrates that it is not plausible that an external attacker could use census tabulations to uncover the characteristics of a particular individual, for three reasons:

- The reconstructed data are usually incorrect.
- The reconstructed data usually do not match even the block, age and sex of anyone identified in outside commercial sources.
- In 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.
  Thus, the system worked exactly as intended. An outside attacker could not use database reconstruction to uncover the characteristics of a particular individual.<sup>7</sup>

Census law mandates that the Census Bureau "shall not make any publication whereby the data furnished by any particular establishment or individual ... can be identified" (Title 13 U.S.C. § 9(a)(2), Public Law 87-813). The Census Bureau's database reconstruction experiment convincingly demonstrates that the 2010 census tabulations meet that standard. The "reconstructed" data is usually false, an intruder would have no means of determining if any inference was true, and an intruder would lack the data needed even to estimate the probability that a re-identification attempt succeeded. Therefore, positive identification of individual respondents by an outsider is impossible, and the data furnished by any particular individual cannot

<sup>&</sup>lt;sup>7</sup> Abowd Decl. App'x B ¶ 24 maintains that in a worst-case scenario (where an external attacker had data that was exactly as accurate and complete as the Census Bureau's internal data) an attacker might be able to guess a respondent's race and ethnicity and be correct in 58% of cases. This statement is false for the reasons I have detailed. It is worth noting, however, that such an exercise would be pointless even if database reconstruction did work as advertised. One could more accurately guess anyone's race and ethnicity just by assigning the most frequent race and ethnic group on the block; that guess would be correct 77.8% of the time. Calculated from U.S. Census Bureau (2011).

be identified. Database reconstruction therefore poses no risk to the Census Bureau's confidentiality guarantee.

# 4. Census disclosure control requires the protection of identities, not concealment of characteristics

The Census Bureau argues that new methods of confidentiality protection are required by census law. The confidentially language in census law first appeared the 1929 Census Act:

No publication shall be made by the Census Office whereby the data furnished by any particular establishment or individual can be identified, nor shall the Director of the Census permit anyone other than the sworn employees to examine the individual reports (Reapportionment Act of 1929, CR 28 § 11).

The current statute is virtually identical, specifying that the Census Bureau "shall not make any publication whereby the data furnished by any *particular establishment or individual ... can be identified*" (Title 13 U.S.C. § 9(a)(2), Public Law 87-813) (emphasis added).

For the past nine decades, the Census Bureau has interpreted the law to mean that Census Bureau publications must protect the identity of respondents. In 2002, this interpretation was codified in the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), which 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" (Title 5 U.S.C. §502 (4), Public Law 107–347).

We have nine decades of precedent, reaffirmed thousands of times by the Census Bureau Disclosure Review Board, reinforcing the interpretation that the Census Bureau is prohibited from publication of statistics that disclose respondent identities. This means that an outsider cannot infer the response of a particular individual, match that response to another database, and have high confidence that the link is correct. The disclosure controls that have been introduced over the past half-century are limited to attributes and circumstances likely to pose a disclosure risk through reidentification. Unlike traditional statistical disclosure control, differential privacy attempts to masking all characteristics, not just individual identities.

The Census Bureau justifies differential privacy through a novel interpretation of census law. According to Abowd (2019), "Re-identification risk is only one part of the Census Bureau's statutory obligation to protect confidentiality. The statute also requires protection against exact attribute disclosure." Under this interpretation, the Census Bureau must not only mask the *identities* of respondents, but also their *characteristics*. Abowd (2019: 16-18), argues in particular that because the 2010 census published the exact number of people of voting age in each census block, that was an exact attribute disclosure and therefore prohibited.<sup>8</sup>

Under this new interpretation, the Census Bureau has been in flagrant violation of the law ever since 1929. Every tabulation of the characteristics of the population necessarily reveals the attributes of individuals. Every census from 1790 to 2010 has published attributes based on exact numbers counted in the census. It is implausible that Congress ever intended to make such exact tabulations of the census illegal.

Differential privacy is oriented to the protection of attributes, not the protection of identities. Accordingly, differential privacy perturbs every attribute tabulated by the census, not just the attributes that pose a risk of enabling re-identification. Because differential privacy focuses

<sup>&</sup>lt;sup>8</sup> The Census Bureau's theory that it is prohibited to disclose the exact number of persons or voting-age persons at the block level is a very recent development. In April 2017 the Census Bureau Disclosure Review Board determined that these counts "can continue to be published as enumerated" (Abowd Decl. App'x B p. 82). When differential privacy was proposed, it specified the publication of exact counts for block population and voting-age populations. According to Garfinkle (2017) and Dajani et. al. (2017), in 2000 the Census Bureau had entered into an agreement with the Department of Justice that required them to publish exact counts of the voting age population of each block. At some subsequent time, the Census Bureau appears to have determined not only that their agreement with the Department of Justice was no longer binding, but that publishing the counts as enumerated was now prohibited.

on concealing individual characteristics instead of protecting respondent identities, it is a blunt and inefficient instrument controlling disclosure of identities.

# 5. The Census Bureau has the capability to return to the 2010 methods for disclosure control

The defendants argue that there is insufficient time to revert to the 2010 standards of disclosure control. Abowd Decl. ¶ 86 states that "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."

That concern appears to be highly exaggerated. The main differences in the system architecture between 2010 and 2020 pertain to the disclosure control software itself; otherwise, the workflow is quite similar (boyd 2020). In both 2010 and 2020, an individual-level data file known as the Census Edited File (CEF) is the input to disclosure control software. The 2020 CEF will be functionally the same as the 2010 CEF. Accordingly, there should be few barriers to applying the 2010 software to the 2020 data. In 2010, the disclosure control software produced an individual-level file known as the Hundred-percent Detail File, or HDF, which was then tabulated to produce PL94-171 and the other census tables. Under differential privacy, the confidentiality protocol is much more complex and takes several steps, but the result is another individual-level file called the Microdata Detail File or MDF, which is then tabulated. In terms of structure, the MDF is equivalent to the HDF, so it is unlikely that using traditional statistical disclosure controls would slow down the tabulation phase. If there are any differences between the HDF and the MDF, the MDF could easily be reformatted.

The defendants raise a second argument that reverting to the 2010 approach to statistical disclosure control would be slow because it would require developing and testing new disclosure

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 19 of 61

control software and procedures. They maintain that it would be impossible to use the 2010 software as it stands, because it provides insufficient disclosure protection. According to Abowd "Knowing that the 2010 Census results were vulnerable to reconstruction, the Census Bureau cannot simply repeat the swapping protocols from the 2010 census," but instead would have to develop an entirely new protocol and software for disclosure protection. *Id.* ¶ 87.

This argument lacks merit, because it is based entirely on the premise that the Census Bureau's database reconstruction experiment showed that the 2010 census was in violation of the confidentiality guarantees governing the census. In fact, however, the database reconstruction showed exactly the opposite. The Census Bureau showed conclusively that even with a massive investment of time, resources, and computing power, it would be impossible for an outside attacker to infer the characteristics of a particular individual respondent from the published tabulations used for the 2010 census.

In sum, the Census Bureau has provided no scientifically credible evidence that reverting to the disclosure control systems used in 2010 would add to the delay in delivery of census results. Given that the most complex remaining aspect of census processing that remains to be completed is the final execution of differential privacy, it is more plausible that substituting the simpler, well-understood protocols from 2010 could actually speed the processing time.

Some computer scientists engaged in privacy research may disagree with my conclusion that database reconstruction poses no threat of re-identification. Everyone would agree, however, that by itself the redistricting file does not enable database reconstruction. As detailed by Abowd Decl. App'x B  $\P$  3, the database reconstruction depends on combining data from nine different tables containing 6.2 billion statistics.

The redistricting file, also known as the PL 94-171 file, provides information on total population and voting-age population broken down by race and Hispanic/Non-Hispanic ethnicity. The Census Bureau database reconstruction experiment relied on additional tables from Summary File 1 (SF1) and Summary File 2 (SF2). These files provide greater detail than is available in the redistricting file, and they are widely used for research and planning. For the 2010 Census, SF1 came out seven months after the redistricting file, and SF2 came out 13 months after the redistricting file.

With respect to the risk of reidentification, the key sensitive tables are the ones that include age distributions: P012 (Sex by Age Group by Block, from SF1), P014 (Sex by Single-year-of-age for the Population under 20 Years by Block, from SF1), and PCT012A-N (Sex by Single-year-of-age by Tract, iterated by Race, from SF2). Without single years of age, there is no possibility for successful re-identification of reconstructed data. The easiest way to prevent attempted database reconstruction would be to withhold one or more of these tables; block-level age detail is seldom needed for research or planning. Alternatively, the Census Bureau could consider additional confidentiality protections for these tables, such as additional swapping or targeted noise infusion on age. Any of these strategies would stymie any chance of re-identification through database reconstruction; other than detailed age, there are no other variables in the census that could enable linkage to an outside source.

The sparse data available in the redistricting file is insufficient to allow database reconstruction; only in combination with other tables can database reconstruction even be attempted. Based on the data production schedules of prior censuses, it is unlikely that the Census Bureau will produce the ancillary tables needed for database reconstruction before the end of 2021, and they may not appear until well into 2022. This means that even if the Census Bureau decides

to implement even stronger protections than existed in 2010 for SF1 and SF2, there is ample time to do so, probably eight months or more.<sup>9</sup>

#### 6. Differential privacy is a poor fit for the protection of census data

There is no guarantee that differential privacy reduces disclosure risk compared with traditional methods of statistical disclosure control. The core metric of privacy loss used in differential privacy is epsilon ( $\varepsilon$ ), which is often referred to as the privacy budget. When  $\varepsilon$  is large, noise infusion is limited and privacy is low, and when  $\varepsilon$  is small, noise infusion is large, and privacy is high. It has long been recognized, however, that there is no direct relationship between the level of  $\varepsilon$  and the risk of disclosing identities. Indeed, McClure and Reiter (2012) demonstrated that the level of  $\varepsilon$  does not determine the level of disclosure risk. Because differential privacy does not target variables and circumstances that are vulnerable to attack, in some datasets with strong differential privacy (low  $\varepsilon$ ), disclosure control can be weak.

According to Hawes and Wright (2021), the Census Bureau is currently planning  $\varepsilon \le 12.3$ for the redistricting data file.<sup>10</sup> This level is far higher than is ordinarily contemplated by privacy researchers. The range of  $\varepsilon$  in the differential privacy literature generally runs from 0.01 to 5.0, but many analysts argue that to guarantee privacy,  $\varepsilon$  should not greatly exceed 1.0 (Lee and Clifton 2011; Dwork 2011). Accordingly, one would expect that  $\varepsilon$ =12.3 would provide a relatively low level of data security.

<sup>&</sup>lt;sup>9</sup> If the Census Bureau strategically applied additional confidentiality protection to tables that include detailed age information, it might create small inconsistencies with the redistricting file, but that would not be a cause for concern; for analysts and planners, maximizing accuracy of the data is far more important than maintaining consistency between the redistricting file and the summary files.

<sup>&</sup>lt;sup>10</sup> Abowd Decl. ¶ 65 indicates that the Census Bureau is planning  $\varepsilon$ =10.3 for persons in its new demonstration product planned for April 2021.

The Census Bureau applied differential privacy to the 2010 Census and re-ran the database reconstruction attack using the differentially private data (LeClerc 2019). The results confirm that differential privacy has limited impact on the vulnerability of the data to reconstruction. With  $\epsilon$ =12.3, LeClerc estimates that approximately 7.5% of the noise-infused population would have "confirmed re-identification" using the same methodology as was employed in the original Census Bureau database reconstruction experiment.<sup>11</sup>

The 7.5% database reconstruction rate reported by LeClerc is lower than the 16.85% rate the Census Bureau reported using only the traditional statistical disclosure control. That does not necessarily mean, however, that the differentially private census files offer greater protection from disclosure of identities than does traditional disclosure control.

The New York Times described a case that effectively illustrates the efficiency of

traditional statistical disclosure control methods:

The bureau has long had procedures to protect respondents' confidentiality. For example, census data from 2010 showed that a single Asian couple — a 63-year-old man and a 58-year-old woman — lived on Liberty Island, at the base of the Statue of Liberty.

That was news to David Luchsinger, who had taken the job as the superintendent for the national monument the year before. On Census Day in 2010, Mr. Luchsinger was 59, and his wife, Debra, was 49. In an interview, they said they had identified as white on the questionnaire, and they were the island's real occupants.

Before releasing its data, the Census Bureau had "swapped" the Luchsingers with another household living in another part of the state, who matched them on some key questions. This mechanism preserved their privacy, and kept summaries like the voting age population of the island correct, but also introduced some uncertainty into the data. (Hanson 2018).

Because the couple lived on a census block with only two residents, the Census Bureau

recognized that they were at high risk of reidentification and thus targeted them for disclosure

<sup>&</sup>lt;sup>11</sup> As Jarmin (2019) acknowledged in his comments on the original database reconstruction experiment, none of these cases could be identified by anyone outside the Census Bureau, so referring to them as "confirmed re-identifications" is misleading.

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 23 of 61

protection. By contrast, differential privacy makes no distinctions between high-risk and low-risk cases, so it infuses noise equally across characteristics and populations. This means that to achieve a given level of disclosure control, differential privacy must introduce far more error than would be needed using traditional statistical disclosure control.

The Census Bureau's database reconstruction exercise does not simulate a realistic attack. We do not know whether realistic attacks, such as the identification of the couple on Liberty Island, would be prevented by differential privacy. Accordingly, based on the information released to date, there is no way to be sure that a differentially private census with  $\varepsilon$ =12.3 will be as secure as a census protected by traditional disclosure controls.

The evidence supports several broad conclusions:

- The statistical disclosure controls employed by the Census Bureau over the past five censuses have proven extraordinarily effective. There is not a single documented case of anyone outside the Census Bureau uncovering the responses of a particular identified person using data from the decennial census.
- The Census Bureau's database reconstruction experiment—the chief rational for adopting differential privacy—failed to demonstrate a credible threat to the exposure of individual identities to anyone outside the Census Bureau. The Acting Director of the Census Bureau confirmed this interpretation when he wrote "The accuracy of the data our researchers obtained from this study is limited, and confirmation of re-identified responses requires access to confidential internal Census Bureau information … an external attacker has no means of confirming them" (Jarmin 2019).
- The Census Bureau's novel contention that census law prohibits "exact disclosure of attributes" even if identities are fully masked is an obvious misinterpretation of the intent of

Congress and contradicts centuries of precedent. Following every census since 1790, the Census Bureau has published exact attributes just as they were enumerated.

- The Census Bureau has not provided evidence that reverting to the disclosure control methods used for the 2010 Census would entail additional delays in the release of redistricting data.
- The statistical tables needed for database reconstruction do not pose the same time constraints as the redistricting data, and there is therefore ample time to implement confidentiality protocols.
- At the proposed privacy budget level, there no guarantee that the Census Bureau's new approach increases protection of identities compared with traditional statistical disclosure controls; in fact, it may provide less protection.

Differential privacy is inappropriate for disclosure control in the census, since it is a blunt and inefficient instrument that adds unnecessary error to every measure, even though most measures pose no risk of a breach of confidentiality (Domingo-Ferrer, Sánchez, and Blanco-Justicia 2020). The adoption of a new regime of disclosure protection is justified only if the benefit of increased protection of respondent identities outweighs the cost inflicted by damage to the integrity of the data. The new method alters the population count for every geographic unit below the state level and changes the counts for every population characteristic. For small populations, the error introduced can easily exceed the true population count. Post-processing for differential privacy also introduces systematic biases in respondent characteristics that can distort the relationships among variables. In short, differential privacy provides little or no documented benefit for the protection of respondent identities, but the costs are devastating.

# References

- Abowd, John. 2017. "Research Data Centers, Reproducible Science, and Confidentiality Protection: The role of the 21st Century Statistical Agency." U.S. Census Bureau. Presentation to the Summer DemSem, June 5, 2017. https://www2.census.gov/cac/sac/meetings/2017-09/role-statistical-agency.pdf
- Abowd, John. 2018a. "How Modern Disclosure Avoidance Methods Could Change the Way Statistical Agencies Operate." Federal Economic Statistics Advisory Committee, December 14 2018. <u>https://www.census.gov/content/dam/Census/about/about-thebureau/adrm/FESAC/meetings/Abowd%20Presentation.pdf</u>
- Abowd, John. 2018b. "Staring-Down the Database Reconstruction Theorem." Presented at the Joint Statistical Meetings, Vancouver, BC, Canada, July 30, 2018

Abowd, John. 2019. Twitter Post, April 7, 2019. https://twitter.com/john\_abowd/status/1114944260837642240

- boyd, dana. 2020. "Balancing Data Utility and Confidentiality in the 2020 US Census." Version of April 27, 2020. <u>https://datasociety.net/wp-content/uploads/2019/12/Differential-Privacy-04\_27\_20.pdf</u>
- Dajani, Aref N. et al. 2017. The modernization of statistical disclosure limitation at the U.S. Census Bureau. <u>https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf</u>
- Domingo-Ferrer, J., Sánchez, D. and Blanco-Justicia, A., 2020. "The limits of differential privacy (and its misuse in data release and machine learning)." arXiv preprint <a href="https://arxiv.org/pdf/2011.02352.pdf">https://arxiv.org/pdf/2011.02352.pdf</a>
- Dwork, C., 2011. "A firm foundation for private data analysis." Communications of the ACM, 54(1), pp.86-95.
- Garfinkel, Simson. 2017. "Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the 2018 End-to-End Test." Presentation at the Census Scientific Advisory Committee, September 17, 2017. <u>https://perma.cc/4J8B-ZEXM</u>
- Garfinkel, Simson, John M Abowd, and Christian Martindale. 2018. "Understanding Database Reconstruction Attacks on Public Data." ACM Queue 16 (5).
- Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. "Issues Encountered Deploying Differential Privacy." In Proceedings of the 2018 Workshop on Privacy in the Electronic Society, 133–137. WPES'18. New York, NY, USA: ACM. <u>https://doi.org/10.1145/3267323.3268949</u>.
- Hanson, Mark. 2018. "To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data." New York Times, Dec. 5, 2018. <u>https://www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html</u>

- Hawes, Michael and Wright, Tommy. 2021. "Summary of DAS Status and P.L. 94-171 Tuning Experiments." Committee on National Statistics, National Academy of Sciences, Expert Groups Webinar, March 30, 2021.
- Jarmin, Ron. 2019. "Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census." Director's Blog, U.S. Census Bureau. <u>https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census\_bureau\_adopts.html</u>
- 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. <u>https://www.census.gov/library/working-papers/2014/adrm/cdar2014-02.html</u>
- Leclerc, Philip. 2019. "The 2020 Decennial Census TopDown Disclosure Limitation Algorithm: A Report on the Current State of the Privacy Loss–Accuracy Trade-Off." In . Washington DC. <u>https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE\_196518</u>.
- Lee, J. and Clifton, C., 2011, October. "How much is enough? choosing ε for differential privacy." In International Conference on Information Security (pp. 325-340). Springer, Berlin, Heidelberg.
- McClure, David and Jerome Reiter. 2012. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." Transactions on Data Privacy, 5: 535-552.
- McKenna, L. 2018. "Disclosure avoidance techniques used for the 1970 through 2010 decennial censuses of population and housing." Technical report 18–47, US Census Bureau, Washington, DC. <u>https://www.census.gov/library/working-papers/2018/adrm/cdar2018-01.html</u>
- McKenna, L. 2019. "U.S. Census Bureau Reidentification Studies." Research and Methodology Directorate, U.S. Census Bureau. <u>https://www.census.gov/content/dam/Census/library/working-papers/2020/adrm/8%20T13%20Reidentification%20studies(tagged)%20CED-DA%20Report%20Series.pdf</u>
- Rastogi Sonia and O'Hara, Amy. 2012. "2010 Census Match Study." 2010 Census Planning Memoranda Series, no. 247. U.S. Census Bureau. <u>https://www.census.gov/content/dam/Census/library/publications/2012/dec/2010\_cpex\_247.p</u> <u>df</u>
- U.S. Census Bureau. 2011. "Census 2010 Summary File1 P5. Hispanic or Latino Origin by Race." Retrieved from <u>https://www.nhgis.org</u>.
- U.S. Census Bureau. 2021. "Census Bureau Administrative Data Inventory March 2021." Retrieved 4/9/2021 from https://www2.census.gov/about/linkage/data-file-inventory.pdf.

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 27 of 61

# Appendix 1

Page 1

# **Steven Ruggles**

Institute for Social Research and Data Innovation 50 Willey Hall, University of Minnesota, Minneapolis, 55455 ruggles@umn.edu, (612) 624-4081

# Education

1985	University of Wisconsin, Demography, Post-Doctoral Trainee
1984	University of Pennsylvania, History, MA, PhD
1978	University of Wisconsin, History, BA

# **Academic Appointments**

2008-	Regents Professor, History and Population Studies, University of Minnesota
2000-	Distinguished McKnight University Professor, University of Minnesota
1995-2008	Professor, University of Minnesota
1988-1995	Associate Professor, University of Minnesota
1985-1988	Assistant Professor, University of Minnesota

# **Administrative Appointments**

Director, Institute for Social Research and Data Innovation, University of Minnesota
Director, IPUMS Center for Data Integration
Director, Minnesota Population Center, University of Minnesota
Director, Social History Research Laboratory, University of Minnesota

# **Elected Offices**

2017-2021	Vice President, President, and Past-President, Social Science History Association
2016-2018	President, Association of Population Centers
2013-2016	President-Elect, President, and Past-President, Population Association of America
2015-2018	Executive Committee, Social Science History Association
2014-2016	Executive Committee, Population Association of America
2001-2003	Secretary, Association of Population Centers
2000-2003	Council member, Inter-University Consortium for Political and Social Research
1999-2002	Executive Committee, Social Science History Association

# **Major National Committees**

2019-	Council on Demographic Data
2016-2018	Advisory Committee on Cyberinfrastructure, National Science Foundation
2015-2016	Chair, Working Group on Open Access to Data in the Social and Behavioral Sciences, National Science Foundation
2010-2015	Advisory Committee for the Social, Behavioral, and Economic Sciences, National Science Foundation
2012-2015	Census Scientific Advisory Committee, U.S. Census Bureau
2012-2014	Chair, Committee of Visitors, Social and Economic Sciences, National Science Foundation
2011-2015	Study Committee on Education for Digital Curation, National Research Council Board on Research Data and Information

2011-2014 Data and Surveys Committee, Social, Behavioral, and Economic Sciences, National Science Foundation

# **Honors and Awards**

2020	Population Association of America Poster Award for "'It's None of their Damn Business': Privacy and Disclosure Control in the U.S. Census, 1790-2020." (with
	Diana L. Magnuson).
2017	Population Association of America Honored Member
	http://www.populationassociation.org/steven-ruggles/
2016	Science Communication and Education Award. Sigma Xi Scientific Research Society.
2014	Named "Wonkblog-Certified Data Wizard." Wonkblog, Washington Post.
2010	Population Association of America Poster Award for "Stem Families and Joint Families in Comparative Historical Perspective."
2009	Warren E. Miller Award for meritorious service to the social sciences. Inter- university Consortium for Political and Social Research.
2009	Platinum Medallion, Delta Airlines.
2005	Scholar of the College, College of Liberal Arts, University of Minnesota.
2003	Robert J. Lapham Award, Population Association of America, for contributions to the application of demographic knowledge to policy issues.
1995	Named "King of Quant." <i>Wired</i> Magazine (March 1995, 86-90). http://archive.wired.com/wired/archive/3.03/ruggles_pr.html
1994	American Historical Review most innovative article on American history published 1992-1994 (AHR nominee for ABC-CLIO America: History and Life Award).
1989	William J. Goode Distinguished Book Award, American Sociological Association, for best book on the family.
1988	Allen Sharlin Memorial Award, Social Science History Association, for best publication in the field of social science history.
1987	McKnight-Land Grant Professorship
1984	National Research Service Award, National Institute of Child Health and Human Development

# **Major Databases**

Integrated Public Use Microdata Series (IPUMS-USA and IPUMS-CPS). This database of over 750 million records provides access to integrated individual-level data from the U.S. decennial censuses of 1790 to 2010, Current Population Surveys from 1962 to the present, and the American Community Surveys from 2000 to the present. A beta-test version of IPUMS appeared in 1993, with subsequent major revisions in 1995, 1998, 2004, 2008, 2014, and 2018. IPUMS has approximately 190,000 registered users and has generated over 20,000 publications. IPUMS was described by the *Journal of American History* as "One of the great archival projects of the past two decades." Liens-Socio, the French portal for the social sciences, gave IPUMS the only "best site" designation that has gone to any non-French website, writing "IPUMS est un projet absolument extraordinaire...époustouflante [mind-blowing]!" <u>http://usa.ipums.org/usa/</u> and <u>https://cps.ipums.org/cps/</u>.

International Integrated Public Use Microdata Series (IPUMS-International). The world's largest population database, IPUMS-International provides information on over a billion persons drawn from 473 censuses and surveys of 102 countries enumerated between 1701 and 2019. A collaboration with 112 national statistical agencies and dozens of scholars around the world, the project has preserved billions of census records, much of it endangered. The first beta release of data for seven countries was in 2002, and there have been annual releases since 2006. Dan Newlon, Director of NSF's Economics Program, remarked that "nothing like this has ever existed anywhere in the world . . . we're now able to move to a Hubble Telescope" (*St. Paul Pioneer Press*, 10/5/04). The project was the sole recipient (among 70 applicants) of NSF's Human and Social Dynamics Infrastructure Award. <u>http://international.ipums.org/international/</u>.

**National Historical Geographic Information System (IPUMS-NHGIS).** This database provides access to all U.S. Census summary data since 1790 and electronic boundary files describing the historical locations of counties and census tracts. The project required gathering and standardizing all surviving aggregate census data from over a million source files; developing comprehensive standardized machine-readable documentation for those data; creating high-precision historical electronic boundary files describing census tracts and counties; and developing web-based tools for disseminating statistical data, geographic data, and metadata. Of the 68 large infrastructure projects in the 2001 NSF infrastructure competition, NHGIS was ranked first, in a category of its own. NHGIS was released in 2007, with 400,000 map polygons, five million lines of tagged and structured metadata, three Terabytes of aggregate data, and a web-based data access system. In 2012, NHGIS began to release *integrated* summary files that provide comparable statistics across census years. <u>http://www.nhgis.org/</u>.

*IPUMS-Terra.* This project, initiated in October 2011, integrates and disseminates global-scale data on population and the environment. In particular, the project makes demographic data interoperable with global environmental data including land cover, land use and climate records. The project began as one of five projects sponsored by the National Science Foundation Office of CyberInfrastucture under the DataNet initiative. The IPUMS-Terra website was named one of GIS Geography's "Top 10 GIS Data Sources" because it is "the next generation resource that can provide researchers or anyone cutting-edge data through validated scientific workflows." <u>GIS Lounge</u> wrote that the project was "at the cutting edge of geospatial information systems" and can "bring people closer together to the information they need to make sense of the world around them." The <u>Map and Geography Libraries Journal</u> gave an article describing IPUMS-Terra its "Best Paper" award for 2016. The first data release was in June 2013. <u>https://terra.ipums.org/</u>

### **Special Journal Issues on Data Infrastructure Projects**

"Big Data." Special issue, Historical Methods, 44:2 (Summer 2011).

"North Atlantic Population Project." Special issue, Historical Methods, 44:1 (Spring 2011).

- "Building Historical Data Infrastructure: New Projects of the Minnesota Population Center." Part 1. Special issue, *Historical Methods*, 36:1 (Winter 2003).
- "Building Historical Data Infrastructure: New Projects of the Minnesota Population Center." Part 2. Special issue, *Historical Methods*, 36:2 (Summer 2003).
- "IPUMS: The Integrated Public Use Microdata Series." Special issue, *Historical Methods*, 32:2 (Summer 1999).
- "The Minnesota Historical Census Projects." Special double issue, *Historical Methods*, 28:1-2 (Winter-Spring 1995).

Page 4

#### Book

Steven Ruggles. 1987. Prolonged Connections: the Rise of the Extended Family in Nineteenth Century England and America Madison: University of Wisconsin Press, xx, 282 pp.

## Articles and chapters

- Steven Ruggles. 2021. The Revival of Quantification: Reflections on Old New Histories." *Social Science History*. 45: 1-25.
- Lara L. Cleveland, Steven Ruggles, and Matthew Sobek. Forthcoming. "Harmonizing Global Census Microdata: IPUMS International." In Irina Tomescu-Dubrow, Christof Wolf, Kazimierz M. Slomczynski, and J. Craig Jenkins (eds) Survey Data Harmonization in the Social Sciences. New York: Wiley.
- Brian Asquith, Brad Hershbein, Tracy Kugler, Shane Reed, Jonathan Schroeder, Steven Ruggles, David Van Riper, and Steve Yesiltepe. Forthcoming. "Assessing the impact of differential privacy on measures of population and racial residential segregation." *Harvard Data Science Review*.
- Steven Ruggles and Diana Magnuson. 2020. "Census Technology, Politics, and Institutional Change, 1790-2020." *Journal of American History* 107: 19-51.
- David Van Riper, Tracy Kugler, and Steven Ruggles. 2020. "Disclosure Avoidance in the Census Bureau's 2010 Demonstration Data Product." In Josep Domingo-Ferrer and Krishnamurty Muralidhar, eds., *Privacy in Statistical Databases*. Berlin and Heidelberg: Springer Verlag, pp. 353-369.
- Steven Ruggles and Diana Magnuson. 2020. "Quantification in History: *Journal of Interdisciplinary History* as a Case Study." *Journal of Interdisciplinary History* 50: 363-382.
- Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. "Differential Privacy and Census Data: Implications for Social and Economic Research." *AEA Papers and Proceedings* 109: 403-408.
- Steven Ruggles, Catherine Fitch, and Evan Roberts. 2018. "Historical Census Record Linkage." Annual Review of Sociology, 44: 19-37.
- Steven Ruggles. 2016. "Marriage, Family Systems, and Economic Opportunity in the United States since 1850." Chapter 1 in Susan M. McHale, Valarie King, Jennifer Van Hook, and Alan Booth (eds). Gender and Couple Relationships. Heidelberg: Springer, pp. 3-41.
- Steven Ruggles. 2016. "Data Sharing in Historical Demography." In Matthijs, K., Hin, S., Kok., J., and Matsuo, H. (eds.) 2016. Upside and down and inside out. The future of historical demography. Leuven: Acco, pp. 99-102.
- Steven Ruggles, Tracy A. Kugler, Catherine A. Fitch, and David C. Van Riper. 2016. "Terra Populus: Integrated Data on Population and Environment." 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 222-231, doi:10.1109/ICDMW.2015.204
- Steven Ruggles. 2015. "Patriarchy, Power, and Pay: The Transformation of American Families, 1800–2015." *Demography*, 52: 1797-1823.
- Peter Baskerville, Lisa Dillon, Kris Inwood, Evan Roberts, Steven Ruggles, Kevin Schurer, and John Robert Warren. 2015. "Mining Microdata: Economic Opportunity and Spatial Mobility in Britain and the United States, 1850-1881." *IEEE BigData 2014*, pp.5-13.
- Robert McCaa, Lara Cleveland, Patricia Kelly-Hall, Steven Ruggles, and Matthew Sobek. 2015. "Statistical coherence of primary schooling in population census microdata: IPUMS-

International integrated samples compared for fifteen African countries." *African Population Studies* 29: 157-180.

- Steven Ruggles, Robert McCaa, Matthew Sobek, and Lara Cleveland. 2015. "The IPUMS Collaboration: Integrating and Disseminating the World's Population Microdata." *Journal of Demographic Economics* 81: 203-216.
- Robert McCaa, Lara Cleveland, Patricia Kelly-Hall, Steven Ruggles, and Matthew Sobek. 2015. "Statistical coherence of primary schooling in IPUMS-International integrated population samples for China, India, Vietnam, and ten other Asia-Pacific countries." *Chinese Journal of Sociology* 1: 333-355. A revised version appeared as Robert McCaa, Lara Cleveland, Patricia Kelly-Hall, Steven Ruggles, and Matthew Sobek. 2016. "相同出生队列小学及以上人口比例统计的一 致性——基于亚太十三国人口普查数据的分析" [The Consistency of the Statistics of the Proportion of the Population of the Primary School and Above in the Same Birth Cohort: An Analysis Based on the Census Data of 13 Asia-Pacific Countries]. *Chinese Journal of Population Science* 1: 53-66.
- Sheela Kennedy and Steven Ruggles. 2014. "Breaking up is Hard to Count: The Rise of Divorce in the United States, 1980-2010." *Demography* 51: 587-598.
- Steven Ruggles. 2014. "Big Microdata for Population Research." Demography 51: 287-297.
- Steven Ruggles. 2012. "The Future of Historical Family Demography." *Annual Review of Sociology* 38: 423–41.
- Lara Cleveland, Robert McCaa, Steven Ruggles, and Matthew Sobek. 2012. "When Excessive Perturbation Goes Wrong and Why: IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata." In Josep Domingo-Ferrer and I. Tinnirello, eds., *Privacy in Statistical Databases*. Berlin and Heidelberg: Springer Verlag, pp. 179-187.
- Steven Ruggles. 2011. "Intergenerational Coresidence and Family Transitions in the United States, 1850-1880." *Journal of Marriage and the Family*, 73: 136-148.
- Steven Ruggles, Matthew Schroeder, Natasha Rivers, J. Trent Alexander, and Todd K. Gardner. 2011. "Frozen Film and FOSDIC Forms: Restoring the 1960 Census of Population." *Historical Methods* 44: 69-78.
- Matthew Sobek, Lara Cleveland, Sarah Flood, Steven Ruggles, and Matthew Schroeder. 2011. "Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center." *Historical Methods* 44: 61-68.
- Steven Ruggles, Evan Roberts, Sula Sarkar, and Matthew Sobek. 2011. "The North Atlantic Population Project: Progress and Prospects." *Historical Methods* 44: 1-6.
- Petra Noble, David Van Riper, Steven Ruggles, Jonathan Schroeder, and Monty Hindman. 2011. "Harmonizing Data across Time and Place: The Integrated Spatio-Temporal Aggregate Data Series." *Historical Methods* 44: 79-85.
- Robert McCaa, Steven Ruggles, and Matthew Sobek. 2010. "IPUMS-International Statistical Disclosure Controls." In Josep Domingo-Ferrer and Emmanouil Magkos, eds., *Privacy in Statistical Databases*. Berlin and Heidelberg: Springer Verlag, pp. 74-84.
- Steven Ruggles. 2010. "Stem Families and Joint Families in Comparative Historical Perspective." *Population and Development Review* 36: 563-577.
- Michael Davern, Steven Ruggles, Tami Swenson, Michael Oakes, and Trent Alexander. 2009. "Drawing Statistical Inferences from Historical Census Data" *Demography* 46: 589-603.

- Steven Ruggles. 2009. "Reconsidering the Northwest European Family System." *Population and Development Review* 35: 249-273.
- Pamela J. Johnson, Lynn A. Blewett, Steven Ruggles, Michael E. Davern, and Miriam L. King. 2008. "Four Decades of Population Health Data: The Integrated Health Interview Series." *Epidemiology*, 19(6):872-875.
- Steven Ruggles and Misty Heggeness. 2008. "Intergenerational Families in Developing Countries." *Population and Development Review* 34: 253-281.
- Steven Ruggles. 2007. "The Decline of Intergenerational Coresidence in the United States, 1850 to 2000" *American Sociological Review* 72: 962-989.
- Steven Ruggles. 2006. "Family and Household Composition" In Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright, eds., *Historical Statistics of the United States: Earliest Times to the Present*. New York and Cambridge: Cambridge University Press, Volume 1: *Population*, pp. 653-689.

Steven Ruggles. 2006. "Linking Historical Censuses: A New Approach." *History and Computing* 14: 213-224.

- Robert McCaa, Albert Esteve, Steven Ruggles, and Matt Sobek. 2006. "Using Integrated Census Microdata for Evidence-Based Policy Making: The IPUMS-International Global Initiative." *African Statistical Journal* 2: 83-100.
- Robert McCaa, Steven Ruggles, Michael Davern, and Tami Swenson. 2006. "IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts." In J. Domingo-Ferrer and L. Franconi (eds.) *Privacy in Statistical Databases*. Berlin and Heidelberg: Springer Verlag, pp. 375-382.
- Steven Ruggles. 2006. "The Minnesota Population Center Data Integration Projects: Challenges of Harmonizing Census Microdata Across Time and Place" 2005 Proceedings of the American Statistical Association, Government Statistics Section. Alexandria, VA: American Statistical Association, pp. 1405-1415.
- Robert McCaa, Steven Ruggles, and Matthew Sobek. 2005. "IPUMS-América Latina: Homologación y diseminación de microdatos censales." *Estudios Demográficos Y Urbanos* 20: 37-70.
- Robert McCaa, Steven Ruggles, and Matthew Sobek. 2005. "Integrated Census Microdata Samples for Scientific Research and Teaching." *Estrela* 8: 25-31. [Japanese]
- Patricia Kelly Hall and Steven Ruggles. 2004. "'Restless in the Midst of their Prosperity': New Evidence of the Internal Migration Patterns of Americans, 1850-1990." *Journal of American History* 91: 829-846.
- Steven Ruggles. 2003. "Multigenerational Families in Nineteenth-Century America." Continuity and Change 18: 139-165. Reprinted in Susan M. Ross, ed., American Families Past and Present: Social Perspectives on Transformations (Rutgers University Press, 2006), Chapter 6.
- Steven Ruggles and Susan Brower. 2003. "The Measurement of Family and Household Composition in the United States, 1850-1999." *Population and Development Review* 29: 73-101.
- Steven Ruggles, Miriam King, Deborah Levison, Robert McCaa, and Matthew Sobek. 2003. "IPUMS-International" *Historical Methods*, 36: 60-65.
- Steven Ruggles, Matthew Sobek, Miriam L. King, Carolyn Liebler, and Catherine Fitch. 2003. "IPUMS Redesign" *Historical Methods* 36: 9-21.

- Catherine Fitch and Steven Ruggles. 2003. "Building the National Historical Geographic Information System." *Historical Methods* 36:1 (2003), 41-60.
- Ron Goeken, Cuong Nguyen, Steven Ruggles, and Walter Sargent. 2003. "The 1880 United States Population Database." *Historical Methods* 36: 27-34.
- Evan Roberts, Steven Ruggles, Lisa Dillon, Ólöf Garðarsdóttir, Jan Oldervoll, Gunnar Thorvaldsen, and Matthew Woollard. 2003. "The North Atlantic Population Project: An Overview." *Historical Methods* 36: 80-88.
- Steven Ruggles. 2003. "The Measurement of Family and Household Composition in Census 2000: An Update." *Population and Development Review* 29: 480-482.
- Steven Ruggles. 2003. "New Projects of the Minnesota Population Center." *Historical Methods*, 36: 5-8.
- Robert McCaa and Steven Ruggles. 2002. "The Census in Global Perspective and the Coming Microdata Revolution." *Scandinavian Population Studies* 13: 7-30.
- Robert McCaa and Steven Ruggles. 2002. "Proyecto Col-IPUMS: Harmonizing the census microdata of Colombia, 1964-2003." In Fernan Vejarano and Robert McCaa, eds., *Homologación de los microdatos censales colombianos. 1964-1993. Memorias del taller Col-Ipums.* Bogotá: DANE.
- Steven Ruggles. 2001. "Living Arrangements and Economic Well-being of the Aged in the Past." *Population Bulletin of the United Nations* 42/43: 111-61.
- Lisa Y. Dillon and Steven Ruggles. 2001. "Creating Historical Snapshots of North America in 1880/1: Collaboration between Historians and the Church of Jesus Christ of Latter-day Saints on the 1880/1 Census Databases of the United States and Canada" *Genealogical Journal* 29: 107-113.
- Steven Ruggles. 2000. "A Data User's Perspective on Confidentiality." Of Significance: A Topical Journal of the Association of Public Data Users 2: 1-5.
- Robert McCaa and Steven Ruggles, "IPUMS International." 2000. In Patricia Kelly Hall, Robert McCaa, and Gunnar Thorvaldsen, eds., *Handbook of International Historical Microdata for Population Research*. Minneapolis: University of Minnesota, pp. 335-346.
- Steven Ruggles, Catherine A. Fitch, Patricia Kelly-Hall and Matthew Sobek. 2000. "Integrated Public Use Microdata for the United States." In Patricia Kelly Hall, Robert McCaa, and Gunnar Thorvaldsen, eds., *Handbook of International Historical Microdata for Population Research*. Minneapolis: University of Minnesota, pp. 259-286.
- Catherine Fitch and Steven Ruggles. 2000. "Historical Trends in Marriage Formation." In Linda Waite, Christine Bachrach, Michelle Hindin, Elizabeth Thomson, and Arland Thornton, eds., *Ties that Bind: Perspectives on Marriage and Cohabitation*. Hawthorne: Aldine de Gruyter, pp. 59-88.
- Steven Ruggles. 1999. "The Limitations of English Family Reconstitution." *Continuity and Change* 14: 105-130.
- J. David Hacker, Steven Ruggles, Andrea R. Foroughi, Brad D. Jarvis and Walter L. Sargent. 1999. "Public Use Microdata Samples of the 1860 and 1870 U.S. Censuses of Population." *Historical Methods* 32: 125-133.
- Todd Gardner, Matthew Sobek, and Steven Ruggles. 1999. "IPUMS Data Extraction System." *Historical Methods* 32: 119-124.

- Matthew Sobek and Steven Ruggles. 1999. "The IPUMS Project: An Update." *Historical Methods* 32: 102-110.
- Steven Ruggles. 1997. "The Rise of Divorce and Separation in the United States, 1880-1990." *Demography* 34: 455-466.
- Steven Ruggles. 1997. "The Effects of AFDC on American Family Structure, 1940-1990." *Journal of Family History* 22: 307-25.
- Steven Ruggles. 1997. "Reply to Oppenheimer and Preston" Demography 34: 475-479.
- Steven Ruggles. 1996. "The Effects of Demographic Change on Multigenerational Family Structure: United States Whites 1880-1980" In Alain Bideau, A. Perrenoud, K.A. Lynch and G. Brunet, eds., *Les systèmes demographiques du passé*. Lyon: Centre Jacques Cartier, pp. 21-40.
- Steven Ruggles. 1996. "Living Arrangements of the Elderly in America, 1880-1980." In Tamara K. Hareven, ed., Aging and Generational Relations over the Life Course: A Historical and Cross-Cultural Perspective. Berlin: Aldine de Gruyter, pp. 254-271.
- Steven Ruggles, Matthew Sobek and Todd Gardner. 1996. "Distributing Large Historical Census Samples on the Internet." *History and Computing* 9: 145-159.
- Steven Ruggles, Matthew Sobek and Todd Gardner. 1996. "Disseminating Historical Census Data on the World Wide Web." *IASSIST Quarterly* 20, 4-18.
- Steven Ruggles and Russell R. Menard. 1995. "The Minnesota Historical Census Projects." *Historical Methods* 28: 6-10.
- Steven Ruggles, J. David Hacker and Matthew Sobek. 1995. "Order out of Chaos: The Integrated Public Use Microdata Series." *Historical Methods* 28: 33-39.
- Steven Ruggles. 1995. "Sample Designs and Sampling Errors in the Integrated Public Use Microdata Series." *Historical Methods* 28: 40-46.
- Steven Ruggles. 1995. "Family Interrelationships." Historical Methods 28: 52-58
- Steven Ruggles. 1994. "The Transformation of American Family Structure." American Historical Review 99: 103-128. Designated best 1994 article on U.S. history in AHR. Ranked 15<sup>th</sup> mostcited AHR article (out of about 2,500 published) as of March 1, 2014, according to Oxford Journals.
- Steven Ruggles. 1994. "The Origins of African-American Family Structure." *American Sociological Review* 59: 136-151.
- Steven Ruggles. 1993. "Historical Demography from the Census: Applications of the American Census Microdata Files." In David Reher and Roger Schofield, eds., Old and New Methods in Historical Demography. New York: Oxford University Press, pp. 383-393.
- Steven Ruggles. 1993. "Confessions of a Microsimulator: Problems in Modeling the Demography of Kinship." *Historical Methods* 26: 161-169.
- Steven Ruggles and Ronald Goeken. 1992. "Race and Multigenerational Family Structure, 1900-1980." In Scott J. South and Stuart E. Tolnay, eds., *The Changing American Family: Sociological and Demographic Perspectives*. Boulder: Westview Press, pp. 15-42.
- Steven Ruggles. 1992. "Migration, Marriage, and Mortality: Correcting Sources of Bias in English Family Reconstitutions." *Population Studies* 46: 507-522.
- Steven Ruggles. 1991. "Comparability of the Public Use Files of the U.S. Census of Population, 1880-1980" Social Science History 15: 123-158.

- Steven Ruggles. 1991. "The U.S. Public Use Microdata Files as a Source for the Study of Long-term Social Change." *IASSIST Quarterly* 15: 20-27.
- Steven Ruggles. 1991. "Integration of the Public Use Files of the U.S. Census of Population, 1880-1980." 1991 Proceedings of the American Statistical Association, Social Statistics Section. Alexandria, VA: American Statistical Association, pp. 365-370.
- Steven Ruggles and Russell R. Menard. 1990. "A Public Use Sample of the 1880 Census of Population." *Historical Methods* 23: 104-115.
- Miriam King and Steven Ruggles. 1990. "American Immigration, Fertility Differentials, and the Ideology of Race Suicide at the Turn of the Century." *Journal of Interdisciplinary History* 20: 347-369.
- Steven Ruggles. 1990. "Family Demography and Family History: Problems and Prospects." *Historical Methods* 23: 22-30.
- Steven Ruggles. 1988. "The Demography of the Unrelated Individual 1900-1950." *Demography* 25: 521-536.
- Susan DeVos and Steven Ruggles. 1987. "The Demography of Kinship and the Life Course." *Life-Span Development and Behavior* 8: 259-281.
- Steven Ruggles. 1986. "Availability of Kin and the Demography of Historical Family Structure." *Historical Methods* 19: 93-102.
- Steven Ruggles. 1983. "Fallen Women: The Inmates of the Magdalen Society Asylum of Philadelphia 1836-1908." *Journal of Social History* 16: 65-82.

#### Reports

- Brief of Amici Curiae Historians to the Supreme Court on the enumeration of undocumented immigrants in the census, 20-366, Donald J. Trump v. New York, Nov. 16, 2020.
- Brief of Amici Curiae Historians in Support of Plaintiffs' Motion for Partial Summary Judgment or Preliminary Injunction. United States District Court, Southern District of New York, 20-cv-05770 (JMF), State of New York v. Donald J. Trump. August 14, 2020.
- Brief of Historians and Social Scientists Margo Anderson, Andrew Beveridge, Rachel Buff, Morgan Kousser, Mae Ngai, and Steven Ruggles as Amici Curiae in Support of Respondents. 18-966, U.S. Dept. of Commerce v. State of New York, April 1, 2019.
- Task Force on Differential Privacy for Census Data. 2018. Implications of Differential Privacy for Census Bureau Data and Scientific Research. MPC Working Paper 2018-6. <u>https://assets.ipums.org/\_files/mpc/MPC-Working-Paper-2018-6.pdf</u>
- Working Group on Open Access to Data in the Social and Behavioral Sciences. 2016. Public Access to NSF-Funded Research Data for the Social, Behavioral, and Economic Sciences. https://www.nsf.gov/sbe/reports/Public\_Access\_NSF\_Workshop\_Report\_Final\_Briefs.pdf
- Jon A. Krosnick, Stanley Presser, Kaye Husbands Fealing, and Steven Ruggles. 2015. *The Future of Survey Research: Challenges and Opportunities*. Report of the Subcommittee on Advancing Survey Research, Social, Behavioral and Economic Sciences. Washington, D.C.: National Science Foundation.
- National Research Council Committee on Future Career Opportunities and Educational Requirements for Digital Curation. 2015. *Preparing the Workforce for Digital Curation*. Washington, D.C: National Academies Press.
- State Budget Trends Commission, Commission Report to the Legislature. January 12, 2009.

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 37 of 61

Ruggles Vita 3/2021

Task Force on Census 2000. The Public Use Microdata Samples of the U.S. Census: Research Applications and Privacy Issues. Prepared by Steven Ruggles with Catherine A. Fitch and Matthew Sobek. Report prepared for the U.S. Census Bureau, Census 2000 Users' Conference on PUMS, May 2000. Washington, D.C.: U.S. Census Bureau.

## **Miscellaneous Publications**

- Steven Ruggles and Matthew Sobek. 2019. "IPUMS." *Encyclopedia of Gerontology and Population Aging*. Springer Nature.
- Steven Ruggles. 2017. "The Importance of Data Curation." Chapter 39 in the Palgrave Handbook of Survey Research, eds, David L. Vannette and Jon A. Krosnick. New York: Palgrave-Macmillan.
- Steven Ruggles. 2017. "Metadata and Preservation." Chapter 71 in the *Palgrave Handbook of Survey Research*, eds, David L. Vannette and Jon A. Krosnick. New York: Palgrave-Macmillan.
- Steven Ruggles. 2015. "Census Microdata." *International Encyclopedia of the Social and Behavioral Sciences*, Second Edition. Oxford: Elsevier, pp. 284-289.
- Steven Ruggles. 2013. "New Data for the Comparative Study of Family and Household" NCFR Report Magazine, Summer Issue, pp. F11-F14.
- Steven Ruggles. 2012. "IPUMS (Integrated Public Use Microdata Series)." In Margo J. Anderson, ed., *Encyclopedia of the U.S. Census*, 2<sup>nd</sup> Edition. Washington, DC: Congressional Quarterly Press.
- Steven Ruggles. 2007. "Integrated Public Use Microdata Series." *International Encyclopedia of the Social Sciences*, 2<sup>nd</sup> Edition. Macmillan Reference USA, Thomson Gale, pp. 382-385.
- Steven Ruggles. 2000. "Living arrangements and Well-Being of the Elderly in the Past." In the *Proceedings of the United Nations Technical Meeting on Population Ageing and Living Arrangements of Older Persons: Critical Issues and Policy Responses*. New York: United Nations.
- Steven Ruggles. 2000. "IPUMS (Integrated Public Use Microdata Series)." In Margo J. Anderson, ed., *Encyclopedia of the U.S. Census*. Washington, DC: Congressional Quarterly Press, pp. 264-267.
- Steven Ruggles. 1999. "Higher Income, Higher Taxes." *Star Tribune*, April 3 1999, p. 17A (Op-Ed).
- Steven Ruggles. 1989. DECOMP, A Program for Multiple Standardization and Demographic Decomposition: Technical Documentation and User's Guide. Minneapolis: Social History Research Laboratory, University of Minnesota. Statistical software with 120 pp. of documentation.

# **Book Reviews**

- Arland Thornton, Reading History Sideways: The Fallacy and Enduring Impact of the Developmental Paradigm on Family Life. Population and Development Review 32 (2006), 174-176.
- Michel Verdon, *Rethinking Households: An Atomistic Perspective on European Living Arrangements. Journal of Family History* 25 (2000), 118-120.
- Stewart Tolnay, The Bottom Rung: African American Family Life on Southern Farms. American Sociological Review 105 (1999), 894-896.

- David Kertzer and Peter Laslett, Aging in the Past: Demography, Society and Old Age. Journal of Interdisciplinary History 27 (1997), 497-498.
- Marguerite Dupree, Family Structure in the Staffordshire Potteries: 1840-1880. American Journal of Sociology 101 (1996), 1437-1439.
- Angelique Janssens, Family and Social Change: The Household as a Process in an Industrializing Community. American Journal of Sociology 100 (1994), 532-533.
- Samuel Preston and Michael Haines, *Fatal Years: Child Mortality in Late Nineteenth-Century America. Journal of Economic Literature* 30 (1992), 2177-2178.
- Elizabeth Pleck, Domestic Tyranny: The Making of American Social Policy against Family Violence From Colonial Times to the Present. Journal of Interdisciplinary History 19 (1989) 686-688.

### **Extramural Funding**

- 2020-2024. Principal Investigator (with Cathy Fitch) RCN: Building an Interdisciplinary Community of Big Microdata Researcher. National Science Foundation SES 2020002, \$499,055.
- 2020-2024 Principal Investigator (with Phyllis Moen) Network for Data-Intensive Research on Aging. P30AG066613. \$904,740.
- 2019-2020 Principal Investigator, "Implications of Differential Privacy on Decennial Census Data Accuracy and Utility." Sloan Foundation, \$124,767.
- 2019-2024 Principal Investigator, "Microdata for Research on Aging in the Global South." R01 AG062601, \$3.1 million), National Institute on Aging. \$3,396,718. Score: 20 (9<sup>th</sup> percentile).
- 2019-2024 Principal Investigator, "International Integrated Microdata Series." SES-1852842, National Science Foundation, \$5,050,000.
- 2018-2023 Principal Investigator, "Microdata for Analysis of Early Life Conditions, Health, and Population." R01AG041831, National Institute on Aging. \$2,889,917. Score: 10 (2<sup>nd</sup> percentile).
- 2018-2023 Principal Investigator, "A Multigenerational Longitudinal Panel for Aging Research." R01AG057679, National Institute on Aging. \$3,396,718. Score: 10 (2<sup>nd</sup> percentile).
- 2018-2023 Co-Investigator, "National Spatiotemporal Population Research Infrastructure." Competing Continuation, R01HD057929, NICHD-DBSB. \$3,250,000. Priority score: 10 (2<sup>nd</sup> percentile)
- 2017-2022 Co-Principal Investigator, "IPUMS Terra: Global Population and Agricultural Data." National Science Foundataion, \$1,473,720.
- 2017-2022 Principal Investigator. "Microdata for Population Dynamics and Health Research." National Institute of Child Health and Human Development, National Institutes of Health. \$3,796,565. Score: 10 (2<sup>nd</sup> percentile).
- 2016-2021 Principal Investigator, "Minnesota Population Center." Competing Continuation, P2C HD041023, NICHD-PDB. \$1,404,530. Priority Score: 10 (no percentile). PI status transferred to John Robert Warren.
- 2016-2021 Co-Investigator, "Integrated Samples of Eurasian Censuses." Competing Continuation, R01HD047283, NICHD-PDB. \$2,994,959. Priority Score: 19 (5<sup>th</sup> percentile).

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 39 of 61

Ruggles Vita 3/2021

- 2015-2020 Principal Investigator, "Big Microdata Expansion Project," with Catherine Fitch and Matthew Sobek. R01HD083829. NICHD-PDB.Total award: \$3,105,210. Priority score: 12 (1st percentile).
- 2015-2020 Co-Investigator, "Models of Demographic and Health Changes following Military Conflict." Principal Investigator: J. David Hacker. R01HD082120, NICHD-PDB.Total award: \$ 2,863,931. Priority score: 21 (11th percentile)
- 2014-2019 Principal Investigator, "International Integrated Microdata Series." National Science Foundation, SES-1357452. Total award: \$6,699,794.
- 2014-2019 Co-Investigator, "Integrated Samples of Latin American Censuses, 1960-2003." Competing Continuation, R01 HD044154, NICHD-DBSB. Principal Investigator: Robert McCaa. \$3,541,813. Score: 10 (5th percentile).
- 2013-2018 Principal Investigator, "Big Data for Population Research," with Catherine Fitch and Matthew Sobek. R01HD078322 National Institute of Child Health and Human Development, Demographic and Behavioral Sciences Branch (NICHD-DBSB).
   \$3,185,694. Priority score: 11 (1<sup>st</sup> percentile)
- 2013-2018 Principal Investigator, "IPUMS Redesign" Competing Continuation, R01 HD43392, NICHD-DBSB. \$2,894,148. Score: 11 (3<sup>rd</sup> percentile).
- 2013-2018 Co-Investigator, "National Spatiotemporal Population Research Infrastructure." Competing Continuation, R01HD057929, NICHD-DBSB. \$3,020,552. Priority score: 11 (3rd percentile)
- 2013-2018 Co-Principal Investigator, "National Historical Geographic Information System." National Science Foundation SES-1324875. \$869,999.
- 2012-2014 Co-Principal Investigator, "Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850-1911." Digging into Data Challenge, National Science Foundation, Economic and Social Research Council (UK), and Social Science and Humanities Research Council (Canada). SMA 1209078, \$120,901 (Minnesota component). Principal Investigator: Evan Roberts.
- 2012-2016 Principal Investigator, "Infrastructure for Population Analysis." (1940 Census Project). National Science Foundation, SES1155572. \$750,000.
- 2012-2017 Principal Investigator, "Baseline Socioeconomic Microdata for Population and Health Research.". R01HD073967, NICHD-DBSB. \$3,043,604. Priority score: 10 (1<sup>st</sup> percentile)
- 2012-2017 Principal Investigator, "Microdata for Analysis of Early Life Conditions, Health, and Population." R01AG041831 National Institute on Aging. \$2,928,170. Priority score: 11 (4<sup>th</sup> percentile).
- 2011-2016 Principal Investigator, "Terra Populus: A Global Population-Environment Data Network." Office of Cyberinfrastructure, National Science Foundation (NSF). \$7,998,550.
- 2011-2016 Co-Investigator, "Integrating, Linking, and Disseminating CPS Data." Competing Continuation, R01 HD047283, NICHD-DBSB. Principal Investigator: John Robert Warren. \$3,111,152. Score: 18 (5<sup>th</sup> percentile).
- 2011-2016 Principal Investigator, "North Atlantic Population Project." R01 HD052110, 2011, \$3,044,475. Score: 14 (3<sup>rd</sup> percentile).

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 40 of 61

- 2011-2016 Principal Investigator, "Minnesota Population Center." Competing Continuation, R24 HD41023, NICHD-DBSB \$1,379,532. Score: 20 (no percentile given).
- 2011-2016 Subaward Principal Investigator, "Early Life Conditions, Survival, and Health: A Pedigree-Based Population Study." University of Utah subcontract (NIA Prime, R01 AG022095), Ken Smith PI. Minnesota component \$669,482.
- 2009-2014 Principal Investigator, "International Integrated Microdata Series." SES-0851414, NSF, \$5,963,296.
- 2009-2014 Co-Investigator, "Baseline Microdata for Analysis of U.S. Demographic Change." R01HD060676, NICHD-DBSB. Principal Investigator: Evan Roberts.\$3,587,689. Score: 120 (2.1 percentile).
- 2009-2014 Co-Investigator, "Integrated Samples of Eurasian Censuses." R01 HD047283, NICHD-DBSB. Principal Investigator: Robert McCaa. \$3,030,118. \$3,587,689. Score: 10 (2nd percentile).
- 2009-2011 Subaward Principal Investigator, "Demographic Data Sharing and Archiving." Subcontract with the Inter-university Consortium for Political and Social Research, Ann Arbor, Michigan; prime funding agency, NICHD-DBSB U24HD048404. \$61,567 (Minnesota component). Score: 150.
- 2009-2014 Co-Investigator, "Integrated Health Interview Series." Competing Continuation, R01 HD046697, NICHD-DBSB. Principal Investigator: Lynn Blewett. \$3,372,067. Score: 137 (11.4 percentile).
- 2009-2012 Co-Principal Investigator, "Minnesota Research Data Center" SES-0851417. NSF. Principal Investigator: Catherine Fitch. \$299,066.
- 2008-2013 Principal Investigator, "Integrated Spatio-Temporal Aggregate Data Series" R01 HD057929, NICHD-DBSB. \$2,919,725. Score: 120 (2.3 percentile).
- 2007-2012 Principal Investigator, "IPUMS Redesign" Competing Continuation, R01 HD43392, NICHD-DBSB. \$3,093,410. Score: 132 (5.8 percentile). Supplement, 2004-2007, \$216,946. Score: 161.
- 2007-2012 Principal Investigator, "New Data Resources from the 1960 Census of Population" R01 HD041575, NICHD-DBSB. \$2,375,491. Score: 120 (1.1 percentile).
- 2007-2012 Principal Investigator, "Public Use Microdata Sample of the 1930 Census." Competing Continuation, R01 HD041575, NICHD-DBSB. \$3,256,088. Score: 126 (2.6 percentile).
- 2007-2012 Co-Investigator, "Integrated Samples of Latin American Censuses, 1960-2003." Competing Continuation, R01 HD044154, NICHD-DBSB. Principal Investigator: Robert McCaa. \$3,100,064. Score: 105 (0.4 percentile).
- 2007-2011 Co-Investigator, "Economic Opportunity and Marriage Formation, 1960-2000" Principal Investigator, Catherine A, Fitch. NICHD-DBSB (R01-HD054643, \$596,505, Score 157, 17 percentile)
- 2007-2010 Co-Principal Investigator, "Economic Opportunity and Marriage Formation, 1960-2000" Co-Principal Investigator, Catherine A, Fitch. NSF Sociology Program (SES-0617560, \$158,706).
- 2007-2010 Co-Investigator, "Population and Health Data Manager." R44 HD053162. NICHD-DBSB. SBIR Grant to Orlin Research, Inc. Principal Investigator, Catherine Ruggles. \$849,237. Score: 194 (no percentile given).

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 41 of 61

Ruggles Vita 3/2021

- 2006-2011 Principal Investigator, "Minnesota Population Center." Competing Continuation, R24 HD41023, NICHD-DBSB, \$1,440,753. Score: 140.
- 2006-2012 Co-Principal Investigator, "National Historical Geographic Information System." BCS-0648005, Geography and Regional Sciences Program, NSF. Co-Principal Investigators: John Adams, William Block, Mark Lindberg, Robert McMaster and Wendy Treadwell. \$1,034,493.
- 2006-2011 Principal Investigator, "North Atlantic Population Project." R01 HD052110, NICHD-DBSB, \$1,398,575. Score: 140 (5.4 percentile).
- 2005-2006 Co-Principal Investigator, "Uncovering the Underrepresented." Grant number 52084, Robert Wood Johnson Foundation. Principal Investigator: Michael Davern. \$350,000.
- 2004-2009 Principal Investigator, "International Integrated Microdata Series." SES-0433654, NSF, Human and Social Dynamics Infrastructure award. \$5,000,000. Sole infrastructure award of 64 applicants to the Human and Social Dynamics initiative.
- 2004-2009 Subaward Principal Investigator, "Demographic Data Sharing and Archiving." Subcontract with the Inter-university Consortium for Political and Social Research, Ann Arbor, Michigan; prime funding agency, NICHD-DBSB U24HD048404. \$437,815 (Minnesota component).
- 2004-2009 Co-Investigator, "Integrated Samples of European Censuses." R01 HD047283, NICHD-DBSB. Principal Investigator: Robert McCaa. \$3,030,118. Score: 120 (1.0 percentile).
- 2004-2009 Co-Investigator, "Integrated Health Interview Series." R01 HD046697, NICHD-DBSB. Principal Investigator: Lynn Blewett. \$2,990,745. Score: 117 (0.3 percentile).
- 2004-2007 Principal Investigator, "IPUMS Redesign" Competing Supplement, R01 HD43392, NICHD-DBSB. \$216,946. Score: 161 (no percentile given).
- 2003-2008 Principal Investigator, "Population Database for the United States in 1880." Competing Continuation. R01 HD39327, NICHD-DBSB. \$2,645,234. Score: 135 (7.2 percentile).
- 2003-2008 Principal Investigator, "Public Use Microdata Sample of the 1900 US Census of Population." R01 HD36451, NICHD-DBSB. \$2,550,000. Score: 120 (2.5 percentile).
- 2003-2006 Co-investigator, "Black Migration to the West." SES-0317254, Sociology Program, NSF; Collaborative research with SES-0317247, University of Washington. Principal Investigators: Stewart Tolnay, J. Trent Alexander, and Jason Digman. \$303,507.
- 2002-2007 Principal Investigator, "IPUMS Redesign." R01 HD43392, NICHD-DBSB. \$2,659,155. Score: 120 (1.8 percentile).
- 2002-2007 Co-Investigator, "Integrated Samples of Latin American Censuses, 1960-2003." R01 HD044154-01, NICHD-DBSB. Principal Investigator: Robert McCaa. \$2,999,934. Score: 120 (1.8 percentile).
- 2002-2007 Principal Investigator, "Public Use Microdata Sample of the 1930 Census." R01 HD041575, NICHD-DBSB. \$3,183,561. Score: 117 (0.4 percentile).
- 2001-2005 Co-Principal Investigator, "National Historical Geographic Information System." BCS-0094908, Geography and Regional Sciences Program, NSF. Co-Principal Investigators: John Adams, William Block, Mark Lindberg, Robert McMaster and Wendy Treadwell. \$4,884,447.
- 2001-2005 Principal Investigator, "Minnesota Population Center." R24 HD41023, NICHD-DBSB, 2001-2011 \$1,327,931. Score: 174.

- 2001-2005 Principal Investigator, "North Atlantic Population Project." SES-0111707, NSF. \$491,506.
- 2000-2001 Principal Investigator, "The 1880 United States Population Database" SES 9910961, Sociology Program, NSF. \$200,000.
- 2000 Principal Investigator, "Microdata Access System." Equipment grant, Sun Microsystems. \$135,000.
- 2000-2003 Principal Investigator, "Population Database for the United States in 1880." R01 HD39327, NICHD-DBSB. \$947,160. Score: 113 (0.3 percentile).
- 2000-2003 Co-Investigator, "Integrated Samples of Colombian Censuses, 1964-2000." R01 HD37508, NICHD-DBSB. Principal Investigator: Robert McCaa. \$549,160. Score: 145 (3.3 percentile)
- 1999-2004 Principal Investigator, "International Integrated Microdata Access System." SBR-9908380, Methodology, Measurement, and Statistics Program, NSF. Co-Principal Investigators: Robert McCaa, Deborah Levison, Todd Gardner, and Matthew Sobek. \$3,501,130.
- 1999-2004 Principal Investigator, "A New Public Use Microdata Sample of the 1910 US Census of Population." R01 HD37888, NICHD-DBSB. \$2,370,000. Score: 131 (7.0 percentile).
- 1998-2003 Principal Investigator, "Public Use Microdata Sample of the 1900 US Census of Population." R01 HD36451, NICHD-DBSB. \$2,100,000. Score: 107 (0.2 percentile).
- 1997-2002 Principal Investigator, "Dissemination & Support of the IPUMS database." SBR-9617820, Methodology, Measurement, and Statistics Program, NSF. \$209,762.
- 1996-2002 Principal Investigator, "Public Use Microdata Samples of the 1860 & 1870 US Censuses." R01 HD34572, NICHD-DBSB. \$1,610,000. Score: 117 (1.8 percentile).
- 1996-2000 Principal Investigator, "Electronic Dissemination of the IPUMS database." R01 HD34714, NICHD-DBSB. \$408,000. Score: 140 (11.7 percentile).
- 1995-1997 Principal Investigator, "Integrated Public Use Microdata Series." Accomplishment-Based Renewal, 1995, SBR-9422805, \$112,130.
- 1994-1996 Subaward Principal Investigator, "Oversample of the 1910 Hispanic Population." (subcontract consortium agreement with University of Texas UT95-0030). R01 HD32325, NICHD-DBSB. \$285,000 (Minnesota component).
- 1993-1998 Principal Investigator, "Public Use Microdata Sample of the 1920 Census." R01 HD29015, NICHD-DBSB. \$2,200,000. Score: 130 (2.2 percentile).
- 1992-1995 Principal Investigator, "Integrated Public Use Microdata Series." SES-9118299, NSF. \$464,913.
- 1992-1994 Co-Principal Investigator, "Public Use Microdata Sample of the 1850 Census." SBR-9210903, Sociology Division, NSF. Co-Principal Investigator: Russell Menard. \$192,203.
- 1989-1993 Principal Investigator, "Public Use Sample of the 1880 Census." R01 HD25839, NICHD-DBSB. Co-Principal Investigator: Russell Menard. \$1,287,000. Score: 125 (1.3 percentile).
- 1984-1985 Principal Investigator, National Research Service Award, National Institute of Child Health and Human Development, National Institutes of Health. \$19,608 (individual postdoctoral award).

Page 16

## University of Minnesota Grants and Awards

- 2008- Regents Professorship, \$50,000 annually.
- 2017- "Minnesota Population Database." \$99,999. Office of the Vice President for Research.
- 2004-2007 Scholar of the College, \$30,000.
- 2000-2004 Distinguished McKnight Professorship. Graduate School, \$100,000.
- 2000-2002 Principal Investigator, "Minnesota Population Center." New Initiatives in Interdisciplinary Research, Graduate School, \$100,000.
- 1994-1995 Principal Investigator, "Public Use Microdata Sample of the 1870 Census: Pilot Study." Grant-In-Aid of Research, Graduate School, University of Minnesota, \$22,000.
- 1992-1993 Principal Investigator, "Public Use Microdata Sample of the 1920 Census: Pilot Study." Grant-In-Aid of Research, Graduate School, University of Minnesota, \$18,000.
- 1990-1991 "Fragmentation of the Family: Living Arrangements in America, 1880-1980." Bush Sabbatical Fellowship, University of Minnesota. \$35,000.
- 1990-1991 Principal Investigator, "Integrated Public Use Microdata Series: Pilot Study." Grant-in-Aid of Research, Graduate School, University of Minnesota, \$15,000.
- 1989-1990 Principal Investigator, "Public Use Sample of the 1880 Census." With Russell R. Menard, Grant-in-Aid of Research, Graduate School, University of Minnesota, \$15,000.
- 1987-1990 McKnight-Land Grant Professorship, Graduate School & McKnight Foundation, \$68,000.
- 1987-1989 Principal Investigator, "The Transformation of American Household Structure, 1880-1980." Three Grants-in-Aid of Research, Graduate School, University of Minnesota, \$30,000.
- 1985-1987 Principal Investigator, "Life-course Transitions and American Family Structure, 1900-1950." Two Grants-in-Aid of Research, Graduate School, University of Minnesota, \$18,000.

# **PhD Advisees Completed**

- William Block, "A Princely Gift Indeed: Agricultural Opportunity, Farm Formation, and Marriage in the United States, 1850-1990" (Ph.D. 2000). Currently Director of the Cornell Institute for Social and Economic Research, Cornell University.
- Lisa Y. Dillon, "Between Generations and Across Borders: Living Arrangements of the Elderly and their Children in Victorian Canada and the United States" (Ph.D. 1997). Currently Professor of Demography, University of Montreal.
- Catherine A. Fitch, "Transitions to Marriage in the United States, 1850-2000" (Ph.D. 2005). Currently Associate Director, Minnesota Population Center.
- Jill Frahm, "Unclaimed Flowers and Blossoms Protected by Thorns: Never-Married Women in the United States, 1880-1930." (Ph.D. 2010). Currently teaches history at the Dakota County Technical College, Rosemount, MN.
- Todd Gardner, "The Metropolitan Fringe: Suburbanization in the United States Before World War II" (Ph.D. 1998). Currently Senior Statistician, U.S. Bureau of the Census.
- Ronald Goeken, "Unmarried Adults and Residential Autonomy: Living Arrangements in the United States, 1880-1990" (Ph.D. 1999). Currently Data Services Core Director and Research Associate, Minnesota Population Center.

- J. David Hacker, "The Human Cost of War: White Population in the Civil War Era" (Ph.D. 1999). Dorothy Thomas Prize, Population Association of America; Finalist for Nevins Prize, Economic History Association. Currently Associate Professor of History, University of Minnesota.
- Patricia Kelly Hall, "Privileged Moves: Migration, Race, and Veteran Status in Post-World War II America" (Ph.D. 2009). Currently Research Associate, Minnesota Population Center.
- Daniel C. Kallgren, "The Individual, the Family and the Community in the Rise of American School Attendance" (Ph.D. 1995). Currently Associate Professor of History, University of Wisconsin Center.
- Diana L. Magnuson, "The Making of a Modern Census: The United States Census of Population, 1790-1940" (Ph.D. 1995). Currently Professor of History, Bethel College, St. Paul, Minnesota.
- Ellen Manovich, "'Is this a Real Neighborhood?': Universities, Urban Development, and Neighborhood Change in the 20<sup>th</sup> Century United States." (Ph.D. 2016).
- Matthew Nelson, "Relieved of These Little Chores: Agricultural Neighbor Labor, Family Labor, and Kinship in the United States 1790-1940." (Ph.D. 2018).
- Evan Roberts, "The Growing Economic Independence of Women": Married Women's Labor Force Participation in the United States, 1860-1940." (Ph.D. 2007). Finalist for the Nevins Prize, Economic History Association. Currently Assistant Professor of Sociology, University of Minnesota.
- David Ryden, "Producing a Peculiar Commodity: Jamaican Sugar Production, Slave Life, and Planter Profits on the Eve of Abolition, 1750-1807." (Ph.D. 1999). Finalist for the Gerschenkron Prize, Economic History Association. Currently Associate Dean and Professor, University of Houston.
- Chad Ronnander, "Many Paths to the Pine: Mdewakanton Dakotas, Fur Traders, Ojibwes, and the United States in Wisconsin's Chippewa Valley, 1815-1837." (Ph.D. 2003). Outreach Director, Hamline University.
- Matt Sobek, "Work in America: Workforce participation and Occupational Attainment in the United States, 1850-1990" (Ph.D. 1997). Currently Data Integration Core Director, Minnesota Population Center, University of Minnesota.

### **Postdoctoral Advisees**

- J. Trent Alexander, Ph.D., History, Carnegie Mellon University. Currently Associate Director, Inter-University Consortium for Political and Social Research.
- Albert Esteve, Ph.D., Demography, Autonomous University of Barcelona. Currently Professor and Director, Center for Demographic Studies, Autonomous University of Barcelona.
- Todd Gardner, Ph.D., History, University of Minnesota. Currently Senior Statistician, U.S. Census Bureau.
- Mark Geiger, Ph.D., History, University of Missouri. Currently Independent Scholar.
- Ronald Goeken, Ph.D., History, University of Minnesota. Currently Data Services Core Director, Minnesota Population Center.
- J. David Hacker. Ph.D., History, University of Minnesota. Currently Associate Professor of History, University of Minnesota.
- Hiromi Ishizawa, Ph.D., Sociology, University of Illinois. Currently Associate Professor of Sociology, Georgetown University.
- Nathan Lauster, Ph.D., Sociology, Brown University; Currently Associate Professor of Sociology, University of British Colombia.
- Carolyn Liebler, Ph.D., Sociology, University of Wisconsin. Currently AssociateProfessor of Sociology, University of Minnesota.
- Berna Torr, Ph.D., Sociology, Brown University. Currently Assistant Professor of Sociology, California State University, Fullerton.

### **Selected Invited Lectures**

Lund University, November 2021

Huber Lecture, the Institute for Population Research, The Ohio State University, April 2020 Center for Demography & Ecology and the Center for Demography of Health and Aging, University of Wisconsin-Madison, March 2020. Population Research Center, University of Texas at Austin, April 2019 Minnesota Population Center Seminar Series, April 2019 University of Pennsylvania, Population Studies Center Colloquim, February 2019 Simons Institute for the Theory of Computing, University of California, Berkeley 2019 University of Pennsylvania, Advanced Demographic Methods Workshop, February 2019 Notestein Seminar, Office of Population Research, Princeton University, November 2018 Autonomous University of Barcelona, Centre d'Estudis Demogràfics, January 2018 University of Minnesota Geography Seminar, Febriary 2018 Center for Studies in Demography and Ecology, University of Washington, October 2017 Minnesota Population Center, March 2017 Colorado University Population Center, November 2016 MITRE Lecture, University of Michigan, March 2016. Sigma XI, University of Minnesota, May 2016. Rand Corporation, Santa Monica, August 2016. Minnesota Population Center, April 2015. Harvard Center for Population and Development Studies, March 2015 Broom Center for Demography, University of California-Santa Barbara, February 2015 California Center for Population Research, UCLA, February 2015 Stanford University, February 2015 Center for Studies in Demography and Ecology, University of Washington, January 2015 Center for Demography and Ecology, University of Wisconsin-Madison, February 2014 Federal University of Minas Gerais, CEDEPLAR, November 2013 Duke University Population Research Institute, November 2013 U.S. Census Bureau, Suitland, Maryland, 2012 Initiative in Population, The Ohio State University, 2012 Center for Demography and Ecology, University of Wisconsin, 2011 Cornell University Population Program, 2011 Institute on the Environment, University of Minnesota, Frontiers Lecture 2011 Max Planck Institute for Demographic Research, Rostock, Germany, April 2010 Notestein Seminar, Office of Population Research, Princeton University, March 2010 Arizona State University Population Program, February 2009 Carolina Population Center, University of North Carolina, November 2008 Max Planck Institute for Demographic Research, Rostock, Germany, July 2008

Population Studies and Training Center, Brown University, November 2007 Maryland Population Research Center, University of Maryland, May 2007 Sociology Colloquium, Stanford University, May 2006 International Institute for Social History, Amsterdam, March 2006 National Science Foundation, Arlington, VA, September 2005 Institut national d'études démographiques, Paris, July 2005 California Center for Population Research, UCLA, May 2005 Center for Studies in Demography and Ecology, University of Washington, April 2004 Department of Demography, University of Montreal, November 2003 Population Studies Center, University of Michigan, 1995 Graduate Group in Demography, SUNY at Buffalo, 1993 Population Studies Center, University of Texas, February 1989 Center for Demography and Ecology, University of Wisconsin, 1985

### **Conference Presentations**

- "Disclosure Avoidance in the Census Bureau's 2010 Demonstration Data Product." With David Van Riper and Tracy Kugler. UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23–25, 2020
- "Big Microdata for Economic History," presented at the Conference on Big Data in Economic History, Institute for Advanced Study, Toulouse. May 28 2020.
- "It's None of their Damn Business': Privacy and Disclosure Control in the U.S. Census, 1790-2020." (with Diana L. Magnuson). Population Association of America, Washington, D.C., April 22-25, 2020.
- "Differential Privacy and Racial Residential Segregation" (with David Van Riper, Tracy Kugler, and Jonathan Schroeder) Presented at the 2020 virtual meeting of the Association for Pimlic Policy Analysis and Management, November 11, 2020.
- "The Revival of Quantification: Reflections on the Old New Histories." Presidential Address, Social Science History Association, Chicago, November 2019.
- "Collaboration of Genealogy and Social Science History: The Case of IPUMS." Social Science History Association, Chicago, November 2019.
- "Building Relationships Where There Are None: Imputing Relationship Status in the 1850, 1860 and 1870 Decennial Census Files." With Jose Pacas and Josiah Grover. Social Science History Association, Chicago, November 2019.
- "Differential Privacy for Population Data," Association of Population Centers, Chicago, October 4, 2019.
- "Assessing the Impact of Differential Privacy on Racial Residential Segregation," Harvard Data Science Symposium, October 25, 2019.
- "The Demography of Kinship: or, My Life as a Microsimulator." Keynote, International Union for the Scientific Study of Population, International Seminar on Kinship and Reproduction in Past Societies. Minneapolis, August 22-23, 2019.
- "Implications of Differential Privacy for Public Data." Keynote Roundtable, Association of Public Data Users, Washington, July 9, 2019.
- "IPUMS Multigenerational Longitudinal Panel." Putting the Pieces Together: Promise, Programs and Pitfalls in Linking Historical and Contemporary Records, Kellog Global Hub, Northwestern University, May 17-19, 2019.
- "Big Census Microdata: IPUMS in the Federal Statistical Research Data Centers," with Catherine Fitch, Erin Meyer, and Todd Gardner. Population Association of America, Austin, April 10-13.

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 47 of 61

- "Imputing Relationship Status in the 1850, 1860 and 1870 Decennial Census Files," with Jose Pacas and Josia Grover. Population Association of America, Austin, April 10-13, 2019.
- "Differential Privacy and Census Data: Implications for Social and Economic Research" Population Association of America, Austin, April 10-13.
- "Differential Privacy and Census Data: Implications for Social and Economic Research." American Economic Association, Atlanta, January 5, 2019.
- "Census Privacy," Privacy in the Digital Era, Institute for Humane Studies, George Mason University. January 3, 2019.
- "Implications of Differential Privacy for Census Bureau Data Dissemination." Federal Economic Statistics Advisory Committee, Washington DC, December 14, 2018
- "Capturing the American People: Census Technology and Institutional Change, 1790-2010." Social Science History Association, Pheonix, November 11, 2018
- "Building a National Longitudinal Research Infrastructure," Family History Technology Workshop, Brigham Young University, Provo, Feb. 27, 2018.
- "Building the Minnesota Population Database." Population Association of America, Denver, April 26-28, 2018.
- "Public Access to Data in the Social, Behavioral, and Economic Sciences." National Science Foundation Workshop on Open Access to Data. Alexandria, VA, Feb. 22.
- "History of the Association of Population Centers." Association of Population Center, Chapel Hill, NC, January 26, 2018.
- "IPUMS Data Integration." INGRID Data Forum on Harmonization and Uses of European Microdata. European Union Research and Innovation Programme, Barcelona, Jan 18 2018.
- "Public Access to Data in the Social, Behavioral, and Economic Sciences." National Science Foundation Workshop on Open Access to Data. Alexandria, VA, Feb. 22
- "Integrating Geographic Context with Individual Situation." With D. Van Riper and J. Schroeder. Social Science History Association, Montreal, November 4, 2017.
- "The National Longitudinal Research Infrastructure." Social Science History Association, Montreal, November 3, 2017.
- "Integrating and Disseminating Large-Scale Microdata." Population Association of America, Chicago, April 27, 2017.
- "Building an National Longitudinal research Infrastructure." Population Association of America, Chicago, April 27, 2017.
- "Building a National Longitudinal Research Structure: Historical Perspectives on Data and Technology," Keynote Address, 2016 FSRDC Annual Research Conference, College Station, TX September 2016.
- "New Spatially Referenced Microdata." 2016 Racial Segregation Conference, College Station, TX September 2016.
- "Data Dissemination and Archiving for the Big Three: An Update." National Academy of Sciences, Annual Meeting of the Standing Committee on the Furture of NSF Supported Surveys, Washington D.C., October 6, 2016.
- "Microdata as a time machine: IPUMS-International population samples illuminate a world we have Gained." With Robert McCaa, Lara Cleveland, Patricia Kelly-Hall, and Matthew Sobek. European Society of Historical Demography conference Leuven, 21-24 September 2016
- "Race Differentials in Marriage, 1960-2013." Population Association of America, Washington, D.C., March 31, 2016.

- "Monitoring Sustainable Development Goals with Data from IPUMS International." Population Association of America, Washington, D.C., March 31, 2016. With R. McCaa, ML King, D. Levison, M. Sobek.
- "Trends in Intergenerational Coresidence in Low and Middle-Income Countries: 1970-2010. With Sheela Kennedy. Population Association of America, Washington, D.C., March 31, 2016.
- "A Note on Data Challenges for the Development Agenda: Observations from IPUMS." United Nations Expert Group Meeting on Strengthening the Demographic Evidence Base for the post-2015 Development Agenda." New York, October 5-6, 2015.
- "Marriageability and the race differential in the frequency of marriage, 1960-2014." Social Science History Association, Baltimore, November 12-15, 2015.
- "The History of Data: Technological Change and the Census, 1790-2020." Plenary lecture, International Association for Social Science Information Services and Technology, Minneapolis, June 2, 2015.
- "Patriarchy, Power and Pay: The Transformation of American Families, 1800-2015." Presidential Address, Population Association of America, San Diego, May 1, 2015.
- "Wage Labor and Family Systems." International Institute of Social History, Amsterdam, December 12-13 2014.
- "The Revolution in Family Formation." Keynote Address, Southern Demographic Association, Memphis, October 17.
- "Complete-Count Data from the U.S. Census." Social Science History Association, Toronto, November 6-9, 2014.
- "Intergenerational mobility in Britain, Canada, and the United States, 1850-1911: New evidence from Digging into Data." With P. Baskerville, L. Dillon, K. Inwood, E. Roberts, K. Schurer. Social Science History Association, Toronto, November 6-9, 2014.
- "Marriage, Family Systems, and Economic Opportunity in the United States Since 1850." Family Symposium, The Pennsylvania State University, October 13-14, 2014.
- "The IPUMS-International partnership enhances the value of census microdata for both producers and users." With R. McCaa, P. Kelly-Hall, L. Cleveland, and M. Sobek. International Association for Official Statistics, Da Nang, Vietnam, October 8-10, 2014.
- "The Decline of Marriage in the United States and Latin America." Population Association of Latin America (ALAP) Lima, Peru, August 12, 2014.
- "Big Microdata from the U.S. Census." With C. Fitch and M. Sobek. Population Association of America, Boston, May 1-3, 2014.
- "Trends in Intergenerational Coresidence in Developing Countries." Poster presentation with S. Kennedy. Population Association of America, Boston, May 1-3, 2014.
- "Terra Populus: Integrated Data on Population and Environment." Coalition for Networked Information, Washington, DC December 8-9, 2014.
- "The Family Transition in the U.S. and Latin America." Family, Gender and Generations: A conference of the Brazilian Population Association. University of Campinas, Nov. 26-27, 2013.
- "Variance Estimation in U.S. Census Data from 1960-2010" Poster presentation with Kathryn Coursolle and Lara Cleveland, Population Association of America, New Orleans, April 11-13 2013.
- "The IPUMS big data revolution: liberating, integrating and disseminating the globe's census microdata." Chaire Quetelet: Demography revisited: The past 50 years, the coming 50 years. Louvain-la-Neuve, November 12-15, 2013.

- "Getting Our Message Across: Strategies and Best Practices to Ensure the Use of Statistics in Decision Making." With Robert McCaa, Lara Cleveland, Albert Esteve, and Matthew Sobek.. International Association of Official Statistics, 12-14 September 2012.
- "Historical Census Microdata and the Future of Household Demography." 2<sup>nd</sup> Mosaic Conference, Budapest, September 6-7 2012.
- "Data Integration, Dissemination, Sustainability, and Metadata." Conference on Surveys and Data, National Science Foundation, November 8-9 2012.
- "The History of Historical Family Demography" Social Science History Association, Vancouver, November 1-4, 2012.
- "Terra Populus: Integrating Data on Population and the Environment." With Catherine A. Fitch, Tracy Kugler, Jonathan Foley, Steven Manson, Matthew Sobek. Population Association of America, San Francisco, May 3-5, 2012.
- "Marriage Formation and Economic Opportunity in the United States: 1970-2000" With Catherine A. Fitch, Sheela Kennedy, and J. Michael Oakes. Population Association of America, San Francisco, May 3-5, 2012.
- "The Future of Historical Family Demography." Keynote address, "Historical Inequality and Mobility: New Perspectives in the Digital Era," University of Guelph, May 25-27, 2012.
- "Terra Populus: A Global Population/Environment Data Network." Presented at Data2012: Coming Together Around Data." Data to Insight Center, Indianapolis, January 25-27, 2012.
- "Marriage and Economic Opportunity in the United States, 1970-2000." With Catherine Fitch and Sheela Kennedy. Social Science History Association, Boston, November 17-20, 2011.
- "The History and Future of Large-Scale Census Data." Keynote Address, Annual Census Research Data Center Conference. Minneapolis, Sept. 15-16, 2011.
- "Terra Populus: A Global Population/Environment Data Network." Workshop on Integrating Global Microdata, 58<sup>th</sup> International Statistical Institute, Dublin, August 20-21, 2011
- "Trends in Divorce and Union Instability in the United States since 1980." With Sheela Kennedy. American Sociological Association, Las Vegas, August 20-23, 2011.
- "Using Restricted-Access Census Data to Study Economic Opportunity and Marriage Formation." With Catherine Fitch and J. Michael Oakes. Population Association of America, Washington, DC, March 31-April 2, 2011.
- "Integrating International Microdata: IPUMS and NAPP." Mosaic Workshop, Max Planck Institute for Demographic Research, Rostock, Germany, May 25, 2011.
- "Measuring Family Interrelationships." Workshop on Poverty Measurement, Urban Institute, Washington, D.C. April 1, 2011.
- "Joint Families and Stem Families and the Northwest European Family System: A Comparative Analysis." Poster presentation, Population Association of America, Dallas, Texas, April 15-17. Winner, 2010 PAA Poster Award.
- "When Comes Baby in the Baby Carriage? Historical Changes in Three Dimensions of Age at Parenthood." With Ann Meier and Catherine Fitch. Population Association of America, Dallas, Texas, April 15-17, 2010.
- "Disseminating Historical Data on the Internet: The IPUMS Experience." International Commission on Historical Demography, 21<sup>st</sup> Congress of Historical Sciences, Amsterdam, August 23-28, 2010.
- "The Decline of Quantiative History." International Commission on Historical Demography, 21<sup>st</sup> Congress of Historical Sciences, Amsterdam, August 23-28, 2010.

- "Historical Census Data for Scientific Research and the Problem of the 1940 Census." National Archives and Records Administration, 1940 Census Workshop. College Park, Maryland, September 13, 2010.
- "Building Dissemination Data Infrastructure." Presented at "Future Investments in Large-Scale Survey Data Dissemination," National Science Foundation, Arlington, VA, July 26-27, 2010.
- "Stem Families and Joint Families in Comparative Historical Perspective." Presented at the plenary session of "The History of Families and Households: Comparative European Dimensions." University of London, 24-26 June 2010.
- "Minnesota Families" Presented at the Annual Conference on Policy Analysis, St. Paul, October 22, 2010.
- "New Directions in Historical Family Demography." Presented at "Nouvelles interrogations en démographie historique et histoire de la famille." 2009 Entretiens du Centre Jacques Cartier, Lyon.
- "Intergenerational Coresidence and Family Transitions in the United States, 1850-1880." Social Science History Association, RMS Queen Mary, Long Beach, CA November 2009.
- "Reconsidering 'Reconsidering the Northwest European Family System'." International Union for the Scientific Study of Population, Marrakech, Morocco, September 28-October 2 2009.
- "Building Social Science Infrastructure." American Sociological Association, San Fransisco, August 7-10 2009.
- "Reconsidering the Northwest European Family System." Population Association of America, Detroit, April 29-May 2 2009.
- "Were Northwest Europe and North America Exceptional? An Analysis of Intergenerational Coresidence." Social Science History Association, Miami, October 23-28, 2008.
- "The Canadian Century Research Infrastructure: International Perspectives." Presented at "State of the World: Information Infrastructure Construction and Dissemination for Humanities and Social Science Research," University of Alberta, Edmonton, October 3-5, 2008.
- "Intergenerational Coresidence and Economic Development: New Evidence from the International Integrated Public Use Microdata Series." (with Misty Heggenness). Presented at "Census Microdata: Findings and Futures," University of Manchester, 1-3 September 2008.
- "IPUMS-International Data Recovery." Presented at "Census Microdata: Findings and Futures," University of Manchester, 1-3 September 2008.
- "Secure Data Laboratories." Workshop on International European Census Microdata, European Association For Population Studies, Barcelona, July 9-12, 2008.
- "Intergenerational Families in Developing Countries." (With Misty Heggeness). Population Association of America, New Orleans, April 16-19 2008.
- "Living Arrangements of the Aged in Comparative Historical Perspective." European Social Science History Conference, Lisbon, February 27, 2008.
- "International Censuses and Intergenerational Families." Plenary Address, Human and Social Dynamics Conference, National Science Foundation. Arlington, VA, October 1-2 2007.
- "Intergenerational Coresidence in Developing Countries: A Comparative Historical Perspective." Social Science History Association, Chicago, November 15-19, 2007.
- "Review of Web-Based Dissemination of the General Social Survey." Workshop on "The General Social Survey: The Next Decade and Beyond," National Science Foundation, Arlington, Virginia, May 2-3, 2007.
- "Using Cyberinfrastructure to Develop Databases for Social Science Research." American Association for the Advancement of Science, San Francisco, February 16-19 2007.

- "IPUMS-International Integrated Census Microdata Extract System: Users and Uses, May 2002-March 2007." (Robert McCaa, Steven Ruggles, and Matt Sobek). 23rd ANCSDAAP Population Census Conference, Christchurch, New Zealand, April 16-18, 2007.
- "Using Census Microdata Disseminated by IPUMS-International to Assess Millennium Development Goals of Literacy, Education and Gender Equity in the Ugandan censuses of 1991 and 2002." (Robert McCaa, Steven Ruggles, and Matt Sobek). Scientific Statistics Conference, Kampala, Uganda, June 11-13, 2007.
- "The Relationship of Socioeconomic Status to Intergenerational Coresidence: A comparative Historical Analysis." (Poster). Population Association of America, New York, March 29-31, 2007.
- "Using Integrated Census Microdata for Evidence-based Policy Making: the IPUMS-International Global Initiative." (Robert McCaa, Albert Esteve, Steven Ruggles, Matt Sobek and Ragui Assaad.) Indian Association for Social Sciences and Health, Third All India Conference, New Delhi, March 16-18, 2006.
- "Disseminating Census Microdata: an Essential Component of National Strategies for the Development of Statistics." (Robert McCaa, Steven Ruggles, and Matt Sobek). Forum on African Statistics Development (FASDEV-II), Addis Ababa, February 6-10, 2006.
- "Archiving Census Microdata: The IPUMS-International Strategy." (Robert McCaa, Steven Ruggles, and Matt Sobek). Forum on African Statistics Development (FASDEV-II), Addis Ababa, February 6-10, 2006.
- "Decline of the Multigenerational Family in the United States." European Social Science History Conference, Amsterdam, March 22-25 2006.
- "The Case for Open Access to Data." Presented at "Disseminating and Analyzing Longitudinal Historical Data," International Institute for Social History, Amsterdam, March 21 2006.
- "Big Social Science History: The Integrated Public Use Microdata Samples." Social Science History Association, Portland, November 3-6 2005.
- "IPUMS International." Human and Social Dynamics Conference, National Science Foundation, Arlington, VA, September 11-12, 2005.
- "Minnesota Population Center Data Integration Projects." Invited paper, session on "Building Historical Data Infrastructure: The Data Integration Projects of the Minnesota Population Center." American Statistical Association, Joint Statistical Meetings, Minneapolis, August 9-11 2005.
- "Intergenerational Coresidence and Economic Opportunity of the Younger Generation in the United States, 1850-2000." Population Association of America, Philadelphia, March 30-April 2, 2005.
- "The Rise of Cohabitation in The United States: New Historical Estimates." With Catherine A. Fitch and Ron Goeken. Population Association of America, Philadelphia, March 30-April 2, 2005.
- "IPUMS-International, IPUMS-USA, and the North Atlantic Population Project: Challenges of harmonizing census microdata across time and place." (Steven Ruggles, Robert McCaa, and Matthew Sobek). Meeting of the International Commission for Historical Demography, Sydney, Australia, July 6, 2005. A revised version of the paper was also presented at the 2005 Biennial meeting of Official Representatives, Ann Arbor, Michigan, October 22, 2005.
- "Are Black Men Marrying Younger than Black Women? New Evidence from Census 2000." (with Catherine A. Fitch). Population Association of America, Boston, April 1-3 2004.
- "The Microdata Revolution: A Brief History." Simposio Homologación y diseminación de microdatos censales. Cartagena, Colombia, January 13-16 2004.

- "IPUMS-International: A Restricted Access Website Providing Anonymized, Integrated Census Microdata for Social Science and Policy Research." (with Robert McCaa, Matt Sobek and Albert Esteve). Invited Paper Meeting 38: Microdata-managing the dilemma between access, privacy, and confidentiality, International Statistical Institute 54th Session, August 15-20, 2003, Berlin.
- "Linked Representative Samples of Nineteenth-Century U.S. Censuses." Social Science History Association, Baltimore, November 13-16 2003.
- "Linking IPUMS Samples to the 1880 Complete Count Census Database." International Microdata Access Group, Conference on "Longitudinal and Cross-Sectional Historical Data: Intersections and Opportunities" Montreal, November 10-11, 2003.
- "Disseminating Anonymized, Integrated Census Microdata via the Internet: the IPUMS-International Project." 20th ANCSDAAP Population Census Conference, Ulaanbaatar, Mongolia, 19-21 June, 2002; co-authors: Robert McCaa and Matt Sobek
- "National Historical Geographic Information System." (with John S. Adams and Catherine A. Fitch). Population Association of America, Atlanta, May 9-11 2002.
- "National Historical Geographic Information System." (with John S. Adams, Robert McMaster and Mark Lindberg). Association of American Geographers, Los Angeles, March 2002.
- "The National Historical Geographic Information System." Social Science History Association, Chicago, November 2001.
- "Proyecto Col-IPUMS: Harmonizing the Census Microdata of Colombia, 1964-2003." (with Robert McCaa). Taller Col-IPUMS: Homologación de los microdatos censales de Colombia, Centro de Investigaciones sobre Dinámica Social (CIDS), Universidad Externado de Colombia. Marzo 23-24, 2001, Bogotá, Colombia.
- "Data Sources for Policy Analysis." Third Upper Midwest Conference on Demography for Policy Makers, St. Paul, November 2001.
- "A Reality Check for IPUMS-International: Labor Force Participation of Mexican Women in Mexico-Census Microdata versus Employment Survey." (with Robert McCaa). Census 2000 and Beyond Conference, Cathie Marsh Centre for Census and Survey Research, University of Manchester, June 22, 2000.
- "Economic Opportunity and Long-term Changes in Age at Marriage in the United States." With Catherine A. Fitch. Social Science History Association, Pittsburgh, October 2000.
- "International Integrated Microdata Access System." IASSIST, Evanston, Illinois, June 2000.
- "The Public Use Microdata Samples of the U.S. Census: Research Applications and Privacy Issues." Census 2000 Users' Conference on PUMS, Alexandria, VA May 22, 2000
- "Economic Opportunity and Marriage Formation in the United States, 1940-1990." With Catherine Fitch. Population Association of America, Los Angeles, March 2000.
- "Living arrangements and Well-Being of the Elderly in the Past." Presented at "Population Ageing and Living Arrangements of Older Persons: Critical Issues and Policy responses." Population Division, United Nations, New York, February 2000.
- "Historical Statistics of the United States: Family Structure." With Susan Brower. Social Science History Association, Fort Worth, November 1999.
- "Moving Through Time: Lifetime Internal Migration Patterns of Americans, 1850-1990." With Patricia Kelly Hall. Social Science History Association, Fort Worth, November 1999.
- "The Decline of the Multigenerational Family in the United States." Presented at "Household and Family in Past Time: New Approaches—New Horizons." University of the Balearic Islands, Palma de Mallorca, Spain, September 1999.

- "The Minnesota Historical Census Project." Presented at "Swedish Population Statistics 250 Years. Comparative Perspectives on the Arrangement and Use of Population Data." Swedish Demographic Database, Umeå University, Sweden, August 1999.
- "Microdata in the Classroom: The IPUMS model." Presented at "2000 and Beyond: Making the Census Accessible." Russell Sage Foundation, New York, September 1998.
- "Demographic Data and Data Dissemination in the New Millennium." Presented at the annual meeting of the Association of Population Centers, Albany, November 1998.
- "The American Family Since 1850." Presented at "Nordisk forskerkurs i historie." University of Tromsø, Norway, June 1998.
- "Accessing Microdata for Public Policy Research." Upper Midwest Conference on Demographics for Policy Makers, Minneapolis, April 1998.
- "Marriage age and Proportions Marrying in the United States, 1850-1880." With Catherine Fitch. Presented at "The Ties that Bind: Marriage in America," a conference of the National Institute for Child Health and Human Development, Bethesda, Maryland, July 1998.
- "Historical Trends in Marriage Formation 1850-1990." With Catherine Fitch. Social Science History Association, November 1998.
- "Global Access to the Integrated Public Use Microdata Series." With Matthew Sobek and Todd Gardner. IASSIST/Computing for the Social Sciences, New Haven, May 1998.
- "The Rise of Divorce and Separation in the United States." Population Association of America, Washington DC, March 1997.
- "The Integrated Public Use Microdata Series." Upper Midwest Conference on Demographics for Policy Makers, Minneapolis, April 1997.
- "The Impact of Welfare on Family Structure." Social Science History Association, New Orleans, October 1996.
- "Electronic Dissemination of Historical Census Data." With Matthew Sobek and Todd Gardner. Population Association of America, New Orleans, May 1996.
- "Disseminating Historical Census Data on the World Wide Web." With Matthew Sobek and Todd Gardner. IASSIST/Computing for the Social Sciences, Minneapolis, May 1996.
- "The Socioeconomic Context of Marital Instability in the United States, 1880-1990." Social Science History Association, Chicago, November 1995.
- "The Integrated Public Use Microdata Series." Biennial meeting of the Inter-University Consortium for Political and Social Research, Ann Arbor, Michigan, October 1995.
- "Making the Integrated Public Use Microdata Series." All-University of California Conference in Economic History on historical public-use microdata, Riverside CA, March 1995.
- "The Effects of Demographic Change on Multigenerational Family Structure." Centre Jacques Cartier Conference, La Plagne, France, December 1994.
- "The Integrated Public Use Microdata Series: Notes on the Preliminary Release." Population Association of America, Miami, May 1994.
- "Race Differentials in Historical Family Structure." Second Carleton Conference on the History of the Family, Ottawa, May 1994. An earlier version of this paper was presented at the Center for Population Policy and Analysis, Humphrey Institute, December 1993.
- "Sources of Bias in Family Reconstitution." Social Science History Association, Baltimore, November 1993.
- "Differential Fertility in 1880." with Miriam L. King. Population Association of America, Denver, 1992.
- "Immigration and Fertility in the Late Nineteenth Century." Social Science History Association, Chicago, November 1992.

- "Integration of the Public Use Files of the U.S. Census." Presented at "Social History: The Challenge of Technology." Economic and Social Research Council, Univ. of Essex, 1991.
- "The Public-Use Census Files as a Source for Social History." Presented at "Old and New Methods in Historical Demography: A Critical Appraisal." International Union for the Scientific Study of Population. Mallorca, Spain, June 1991.
- "Proposal for an Integrated Census Microdata Series." International Association for Social Science Information Service and Technology, Edmonton, Alberta, May 1991.
- "Living Arrangements of the Elderly in America, 1880-1980." Social Science History Association, New Orleans, October 1991.
- "Old Age and Multigenerational Family Structure Since 1880." Presented at "Aging and Generational Relations." National Institutes of Health, Center for Family Research, University of Delaware, October 1991. An earlier version of this paper was presented at the Cambridge (University) Ageing Seminar, June 1990.
- "Comparability of the Historical Public Use Samples." American Statistical Association, Atlanta, August 1991.
- "Migration, Marriage, and Mortality: Correcting Sources of Bias in English Family Reconstitutions." Presented at the Cambridge Group for the History of Population and Social Structure, May 1990.
- "Race and Multigenerational Family Structure, 1900-1980." Presented at "Demographic Perspectives on the American Family: Patterns and Prospects." The Albany Conference, SUNY-Albany, April 1990.
- "The Belated Decline of the Extended Family." American Association for the Advancement of Science, New Orleans, February 1990.
- "Immigration, American Fertility Differentials, and the Ideology of Race Suicide." With Miriam King. Population Association of America, New Orleans, April 1988. Various other versions of this paper were presented at the Social Science History Workshop, University of Texas, Austin, February 1989, and appeared as Center for Demography and Ecology *Working Papers* 85-13.
- "Fertility and Marriage Among Second-Generation Immigrants in the Late-Nineteenth Century." Social Science History Association, New Orleans, November 1987.
- "Confessions of a Microsimulator: Problems in Modeling the Demography of Kinship." Presented at a conference on demographic microsimulation sponsored by the International Institute for Applied Systems Analysis, Budapest, Hungary, November 1987
- "The Use and Misuse of Simulation for the Historical Study of Family Structure." Presented at "New Directions for Demographic History: A French-American Roundtable," a Sloan Foundation conference organized by Charles Tilly and Olivier Zunz. New York, November 1986.
- "Aging and Family Structure: An Historical Perspective." Wisconsin Congress on Aging, Madison, Wisconsin, January 1985.
- "Kinship and the Life Course." With Susan De Vos. Am. Sociological Association, Washington, D.C., August 1985. Center for Demography and Ecology *Working Papers* 85-15.
- "Simulation and the Measurement of Historical Family Structure." Population Association of America, Boston, March 1985. A version of this paper, "The Use of Standard Propensities for the Historical Analysis of Extended Family Structure." Center for Demography and Ecology *Working Papers* 85-14.
- "Non-Traditional Family Structures in Post-War America." Testimony presented before the Wisconsin Equal Opportunities Commission, Task Force on Alternative Families, Madison, Wisconsin, November 1984.

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 55 of 61

Ruggles Vita 3/2021

"Microsimulation and its Application to the Historical Study of Family Structure." Social Science History Association, Nashville, October 1981.

### **Commenter/Panelist**

- Discussant, "Comparative Family Perspectives" Population Association of America, Washington, April 22-25, 2020
- Organizer and discussant, "Controversies of Counting: Race, Citizenship, and Privacy in the Census." American Sociological Association, August 9, 2020.
- Organizer and Discussant, "Automatic Handwriting Recognition." Social Science History Association, Chicago, November 2019.
- Discussant, "Large Linkage Projects: New Opportunities." Social Science History Association, Phoenix, November 9, 2018.
- Chair/Organizer, "New Findings on Neighborhoods and Mobility From the Complete Count Census Microdata" Population Association of America, Denver, April 26-28, 2017.
- Chair/Discussant, "Residential and Social Mobility in the United States: New Studies Using IPUMS." Social Science History Association, Montreal, November 2, 2017.
- Chair/Discussant, "Presidential Session: Interdisciplinarity in the Big Data Age." Social Science History Association, Chicago, November 17-20, 2016.
- Discussant, "Author Meets Critics: Tolnay and Bailey, Lynched: The Victims of Southern Mob Violence." Social Science History Association, Chicago, November 17-20, 2016.
- Discussant, "American Community Survey Content Review," U.S. Census Bureau, Suitland, MD, April 2015.
- Discussant, "Privacy and Confidentiality," National Science Foundation, Arlington, October 2014.
- Discussant, "Electronic Data Collection," U.S. Census Bureau, Suitland, MD, April 2014.
- Chair/Discussant, "Spatiotemporal Demographic Analysis," Social Science History Association, Chicago November 22, 2013.
- Chair/Panelist, "Talk Data to Me: A Conversation with Historians about Using Large-Scale Digital Data in Research and Teaching." American Historical Association, Chicago, January 5-8, 2012.
- Panelist, "New Developments in Data Digitization Projects." 2<sup>nd</sup> Mosaic Conference, Budapest, September 6-7 2012.
- Chair, "Spatial Variation in Residence Patterns." Social Science History Association, Boston, November 17-20, 2011.
- Panelist, "Roundtable Discussion: Daniel Scott Smith: A Tribute to and Critical Reappraisal of his Scholarship." Social Science History Association, Chicago, November 18-21, 2010.
- Discussant, "Typologies of Families and Households: New and Old Approaches." Social Science History Association, Chicago, November 18-21, 2010.
- Panelist, "Forging the Future for Sociological Research: Building Infrastructure for Disciplinary, Interdisciplinary and Multi-disciplinary Research" American Sociological Association, San Francisco, August 8, 2009.
- Panelist, "Approaches and Methods for the Study of Individuals, Families, and Households." Social Science History Association, Miami, October 23-26, 2008.
- Chair, "Family Change in Historical Perspective." Population Association of America, New Orleans, April 16-19 2007.
- Chair, "Historical Demography." Population Association of America, New York, March 30-April 1 2007.

- Panelist, "Global Family Theory: Past, Present, and Future." Theory Construction and Research Methodology Plenary Session, National Council on Family Relations. Minneapolis, November 7, 2006.
- Panelist, "At the Shrine of the Bitch Goddess: The Future of Quantitative History." Social Science History Association, Minneapolis, November 3, 2006.
- Chair, "Long-term Change in Family Life-Course Transitions." Population Association of America, March 30, 2006.
- Panelist, "Author Meets Critics: Arland Thornton, Reading History Sideways: The Fallacy and Enduring Impact of the Developmental Paradigm on Family Life." Social Science History Association, Portland, November 3-6 2005.
- Panelist, "Big Social Science History: Big Cities, Big Histories—Megaresults For Megabucks? Retrospective And Prospective Looks At Large-Scale Quantitative Urban History Projects." Social Science History Association, Portland, November 3-6 2005.
- Panelist, "Roundtable: National Variations in Historical Microdata Projects." Social Science History Association, Chicago, November 18-21 2004.
- Chair/Discussant, "The Organization and Composition of Households across Time and Space." Social Science History Association, Chicago, November 18-21 2004.
- Panelist, "Public Needs and Private Information." Census Bureau Symposium, "America's Scorecard: The Historical Role of the Census in an Ever-Changing Nation." Woodrow Wilson International Center for Scholars, Washington D.C., March 4-5 2004.
- Chair, "The Family in Historical Perspective." Population Association of America, Boston, April 1-3, 2004.
- Chair, "The Encyclopedia of Population: Concepts, Consensus, Conflict." Social Science History Association, , Baltimore, November 13-16 2003.
- Chair, "Findings from U.S. Historical Censuses." Population Association of America, Minneapolis, May 1-3, 2003.
- Chair, "Geographic Information Systems II." International Association for History and Computing, Tromsø, Norway
- Chair, "North Atlantic Population Project: Methods and Prospects." Social Science History Association, St. Louis, October 2002.

Chair, "Census 2000: New Trends." Population Association of America, Atlanta, May 9-11, 2002.

Chair, "Who Counts? The Politics of Census Taking in Contemporary America." Keynote session, Third Upper Midwest Conference on Demography for Policy Makers, St. Paul, November 2001.

- Panelist, "The Role of Location in Social Science History: Is There a 'Spatial Turn'?" Social Science History Association, Chicago, November 2001.
- Chair, "Census 2000: An Unparalleled Research Resource." 25<sup>th</sup> Biennial Meeting of Inter-University Consortium for Political and Social Research Official Representatives. Ann Arbor, Michigan, October 2001.
- Panelist, "Conference on Data Dissemination and Archiving." Demographic and Behavioral Sciences Branch, National Institute of Child Health and Human Development. Rockville, Maryland, October 2001.
- Chair, "Retrospective: Census Accuracy, Past and Present." Social Science History Association, Pittsburgh, October 2000.
- Chair/Discussant, "Long-Term Changes In Household and Family Structure." Social Science History Association, Pittsburgh, October 2000.
- Chair, "Historical Demography." Population Association of America, Los Angeles, March 2000

- Chair/Panelist, "The Future of Social Science History Computing in The New Millennium: A Roundtable Discussion." Social Science History Association, Fort Worth, Nov. 1999.
- Panelist, "The Future Of SSHA Program Planning On The Web." Social Science History Association, Fort Worth, November 1999.
- Rapporteur on operational issues and new technology at "Social Sciences for a Digital World: Infrastructure Needs of the Social Sciences." OECD conference, Ottawa, Oct. 6-8 1999.
- Discussant, "Methods and Models in Historical Demography." Population Association of America, Chicago, March 1998.
- Panelist, "Authors meet Critics: The Cambridge Group Family Reconstitutions." Social Science History Association, Washington DC, November 1997.
- Discussant, "Canadian Families Project" Social Science History Association, Washington DC, November 1997.
- Panelist, "Roundtable on Peter Laslett." Social Science History Association, New Orleans, October 1996.
- Discussant, "Migration and the Life Course in Comparative Perspective." Social Science History Association, Atlanta, November 1994.
- Discussant, "Studies in Family Structure and Kinship." Second Carleton Conference on the History of the Family, Ottawa, May 1994.
- Discussant, "The Problem of Migration in Family Reconstitution." Social Science History Association, Baltimore, November 1993.
- Discussant, "Studies in the History of Class and Ethnicity." Social Science History Association, Chicago, November 1992.
- Discussant, "Matrilineality and Patrilineality in Historical Perspective." University of Minnesota Conference on Matrilineality and Patrilineality, Minneapolis, May 1992.
- Discussant, "Historical Change in Aging and Life Course Patterns." Population Association of America, Washington, D.C., March 1991.
- Discussant, "Public Use Sample of the 1880 Census." Social Science History Association, November 1989.
- Discussant, "Single Parenthood in 19<sup>th</sup> Century Brazil and the United States." Social Science History Association, November 1989.
- Discussant, "New Approaches to the Measurement of Household Structure and the Life Course." Social Science History Association, November 1988.
- Discussant, "The European Peasant Family and Economy." University of Minnesota Conference on Peasant Families, Minneapolis, October 1988.
- Discussant, "Living Arrangements of the Elderly in Preindustrial Populations." Social Science History Association, Chicago, November 1985.

### National and International Professional Service

Board of Directors, Association of Population Centers, 2000-

- Selection Committeee, Miller and Flanigan Awards, Inter-University Consortium for Political and Social Research, 2019.
- Nominator and presenter, Laureate Award, International Union for the Scientifc Study of Population 2018
- Nominator and presenter, Warren Miller Award, Inter-University Consortium for Political and Social Research, 2017.
- Performance Monitoring and Accountability 2020 (PMA2020) Consutative Group, 2016-2017 Population Association of America Advocacy Days, 2015, 2017

Page 31

Board of Management, Mosaic Project, Max Planck Institute for Demographic Research, 2012- 2016
Committee on Government and Public Affairs, Association of Population Centers, 2017-2018
Development Committee, Population Association of America, 2016
Co-Chair, Program Committee, Population Association of America, 2014-2015.
Initiatives Committee, Population Association of America, 2016
Executive Committee, Population Association of America, 2014-2016
Committee on Finance, Population Association of America, 2014-2016
Committee on Government and Public Affairs, Population Association of America, 2014-2016
International Advisory Board, Integrated Census Microdata Project (U.K.), 2009-2014
Award Committee, Inter-University Consortium for Political and Social Research, 2010-2011.
Chair, International Advisory Board, "Life Courses in Context." International Institute of Social History, Amsterdam (member, 2003-2011; chair, 2006-2011)
Coalition for National Science Funding, 2011.
Planning Committee, Workshop on Social Observatories, National Science Foundation, 2010.
Member, State Budget Trends Commission, State of Minnesota, 2007-2009
Co-Chair, Robert J. Lapham Award Committee, Population Association of America, 2006
Representative, Coalition for National Science Funding. Courtesy visits with members and staff of the House of Representatives and Senate, September 13-14, 2005
Nominator and presenter, Warren Miller Award, Inter-University Consortium for Political and Social Research, 2005.
Chair, Allen Sharlin Award Committee, Social Science History Association (member, 2003-2005; chair, 2004-2005)
Chair, Census 2000 Committee, Inter-University Consortium for Political and Social Research, 2000-2004
Chair, Nominating Committee, Association of Population Centers, 2004
Chair, Archival Development Committee, Inter-University Consortium for Political and Social Research. 2002-2003
Nominating Committee, Inter-University Consortium for Political and Social Research, 2002-2003, 2009
Information Technology Committee, Inter-University Consortium for Political and Social Research, 2000-2003
Chair, Task Force on Census 2000, 1999-2000.
Chair, Family and Demography Network, Social Science History Association, 1996-1997
Technical Advisory Group, Immigration and Naturalization Service Backfile Rescue, 1996-1997

### **Editorial Boards**

Demography, 2010-present Journal of Interdisciplinary History, 2002-present Historical Life-Course Studies, 2013-present Social Science History, 2001-2013 H-Demog (H-Net) 1996-2000 Historical Methods, 1996-2001 Historical Statistics of the United States, 1996-2005

### Editing

Co-Editor, Social Science History, 1991-1996.

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-6 Filed 04/20/21 Page 59 of 61

Ruggles Vita 3/2021

Page 32

#### **Peer reviewer of Grant Proposals**

Austrian Science Fund (FWF); BD2K Solicitation, National Institutes of Health; Canada Foundation for Innovation; Digital Libraries Initiative Panel (National Science Foundation, National Endowment for the Humanities, and Department of Defense); Dutch Council for the Humanities; Economic and Social Research Council (UK); MacArthur Fellowship; National Institute on Aging, program project review panels (2007, 2009, 2011, chair 2013); National Institute on Aging, special panel on population aging research centers (2004, 2009, 2014); National Institute of Child Health and Human Development, Population Research Infrastructure Program Study Section; National Institute of Child Health and Human Development, program project review panel (chair, 2013); National Institutes of Health, Social Sciences and Population Studies ZRG1 HOP-B 90 S (2006, 2008); National Institutes of Health; National Science Foundation, reviews for Sociology, Economics, Measurement Methods and Statistics, and several interdisciplinary solicitations; Research Council of Katholieke Universiteit Leuven; Riksbankens Jubileumsfond, Sweden; Royal Netherlands Academy of Arts and Sciences (KNAW); Russell Sage Foundation; Nederlandse Organiisatie voor Wetenschappelijk Onderzoek (NWO); Social Sciences and Humanities Research Council (Canada); Swedish Council for Planning and Coordination of Research; Swedish Research Council; Swiss National Science Foundation; Wellcome Trust (UK)

#### Peer Reviewer of Book Manuscripts

Cambridge University Press, Houghton-Mifflin, Kluwer Academic Publishers, MIT Press, National Academies Press, Russell Sage Press, University Press of Virginia

#### **Peer Reviewer of Articles**

Ageing & Society; American Journal of Sociology; American Sociological Review; Cahiers Québécois de démographie; Caribbean Studies; Comparative Studies in History and Society; Continuity and Change; Demographic Research; Demography; Economic Geography; European Journal of Population; European Review of Economic History; Gender & History; Harvard Review of Psychiatry: Historical Life-Course Studies: Historical Methods: International Journal of Epidemiology; International Journal of Health Geographics; International Migration Review; Journal of American History; Journal of Business and Economic Statistics; Journal of Gerontology: Social Sciences; Journal of Family History; Journal of Family Issues; Journal of Interdisciplinary History; Journal of Marriage and the Family; Journal of Official Statistics; Journal of Population Research; Journal of Sustainable Development; Journal of the American *Medical Informatics Association; PLOS ONE; Population; Population and Development Review;* Population Research and Policy Review; Population, Space, and Place; Population Studies; Privacy in Statistical Databases; Proceedings of the National Academy of Sciences; Professional Geographer; Religions; Research on Aging; Social Forces; Social Problems; Social Science History; Social Science Research; Social Sciences; Sociological Forum; Sociological Perspectives; Sociological Quarterly; Socius; Statistical Journal of the IAOS; The History of the Family: An International Quarterly; William and Mary Quarterly

### **External Reviewer for Promotion and Tenure**

Binghamton University, Brown University, University of California-Los Angeles, Catalan Ministry of Universities, City University of New York, Max Planck Institute, Oklahoma State University, University of Colorado, University of Houston, University of Maryland, University of Michigan, Stanford University, University of Washington, University of Utah

### **Selected Conference Organizing**

Social Science History Association, 2019.

Page 33

Population Association of America, Program Co-Chair, 2014-2015

Population Association of America: Invited Session Organizer, ten sessions, 1999-present

- Social Science History Association: Program Chair, 1998-1999; Program Committee, 1996-1999; Session Organizer, 50+ conference sessions, 1988-present
- Life Course Center, University of Minnesota: Program Committee, "Changing Lives and Changing Times: American Life Courses in Historical Perspective" (Conference, Minneapolis 2007)
- International Union for the Scientific Study of Population: Program Committee, "Space and Time in Historical Demographic Studies," 2005-2006
- Upper Midwest Conference on Demography for Policy Makers: Organizing Committee, 1997, 1998, 2001.

### **Search Committees**

Faculty Search in Statistical Sociology, 2018-2019
Vice President for Research, University of Minnesota, 2017-2018
Faculty search for environmental history, Department of History, 2016-2017
Division Head, Health Policy and Management, 2015-2016
Faculty search, Health Policy and Management, School of Public Health, 2006-2007
Associate Vice President for Research Administration, University of Minnesota, 2006-2007
Chair, Search Committee for Rudolph J. Vecoli Chair in Immigration History, 2004-2005
Faculty search in American History, Department of History, 1999-2000
Faculty search, Center for Population Policy and Analysis, Humphrey Institute, 1992
Faculty search in women's history, Department of History, 1990-1991

### Service to the University of Minnesota

Administrative Council, Office of the Vice President for Research 2019-Task Force on Data Science, College of Science and Engineering 2019-Research Computing Internal Review Committee, 2018-2019 Grand Challenges Research Strategies Group, 2015-2016 Health Informatics Steering Committee, University of Minnesota, 2011-2015 Information Exchange Executive Oversight Committee, University of Minnesota, 2010-2015 University of Minnesota Interdisciplinary Informatics Executive Steering Committee 2011-2015 McKnight Distinguished University Professor Committee 2011-2016 Health Informatics Steering Committee, 2011-2018 U-Spatial Advisory Board, 2011-present Minnesota Research Data Center Steering Committee, 2011-present Chair, Data Governance and Security Committee, Academic Health Center, 2010-2014 Faculty Senate Committee on Committees, 2010-2012 Interdisciplinary Research Centers Working Group, 2008-2011 University Metropolitan Consortium, 2007-2011 Provost's Interdisciplinary Advisory Group, 2007-2011 Chair, Research Committee, University Senate, (member, 2004-2009; chair, 2005-2007) Joint Subcommittee on Faculty Activity Databases, Faculty Senate, 2007-2008 Provost's Research Council, University of Minnesota, 2005-2007 Faculty Consultative Committee, University of Minnesota, 2005-2007 Faculty Senate (ex officio), University of Minnesota, 2005-2007 Task Force on Collaborative Research, University of Minnesota, 2005-2006

Page 34

Graduate School Fellowship Committee, 2002-2005 Faculty Summer Research Review Committee, Graduate School, 1999-2001 Fulbright Fellowship Nominating Committee, University of Minnesota, 1997-1998 West Bank Union Board of Governors, University of Minnesota, 1987-1988

### Service to the College of Liberal Arts

Interdisciplinary Collaborative Workshop Review Committee, 2017-2018 Social Sciences Working Group, 2016-CLA Research Roadmap Group, 2014 Tenured Hire Review Committee, 2006-2012 Promotion and Tenure Committee, College of Liberal Arts, 1998-1999 Social Science Research Facilities Advisory Committee, College of Liberal Arts, 1994-2003 Information Technology Committee, College of Liberal Arts, 1994-1997 College of Liberal Arts Computing Facilities Committee, 1991-1992 Course review committee, College of Liberal Arts, 1990-1991

### Service to Department of History

Promotion, Tenure, and Merit Committee, Department of History, 2013-2015 (chair), 2010-2012, 2007-2009 (co-chair), 2003-2005, 1999-2000, 1996-1998 (chair), 1993-1994, 1990-1992, 1987-1988 Merit Committee, Department of History (elected), 2016-2018. Promotion and Tenure Committee, Department of History (elected), 2018-2020 History Department Advisory Committee (elected), 2015-2016, 2020. Leadership Group, Carnegie Initiative on the Doctorate, Department of History, 2003-2005 U.S. Field Screener, Graduate Studies, Department of History, 1997-2000, 2002-2004 Graduate Studies Committee, Department of History, 2001-2003, 1998-1999, 1993-1994 Faculty Awards and Prizes Committee, Department of History, 2000-2001 Director, Social History Research Laboratory, Department of History, 1990-1998 Faculty Advisor, Undergraduate History Association, Department of History, 1987-1989 Director of Undergraduate Studies and Chair of Undergraduate Studies Committee, Department of History, 1986-1989 Advisory Committee, Department of History, 1986-1989, 2020-Computer Use Officer, Department of History, 1985-1989 Curriculum Committee, Department of History, 1985-1989

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-7 Filed 04/20/21 Page 1 of 2

# Exhibit 7

### Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census

February 15, 2019 WRITTEN BY: DR. RON JARMIN, DEPUTY DIRECTOR AND COO

The Census Bureau takes seriously its legal and professional obligation to safeguard the information it gathers from the public. For this reason, we are modernizing and strengthening how we protect privacy in the statistics we release starting with the 2020 Census. We are deploying differential privacy, the gold standard for privacy protection in computer science and cryptography, to preserve confidentiality in the 2020 Census and beyond. Differential privacy was developed by researchers at Microsoft and is now utilized by many leading tech firms. There are many variants of differential privacy. The one selected for the 2020 Census introduces controlled noise into the data in a manner that preserves the accuracy at higher levels of geography.

Since the last Decennial Census, the data world has changed dramatically. Much more personal information is available online and from commercial providers, and the technology to manipulate that data is more powerful than ever. Because we are sworn by law to protect your data, we are constantly testing and improving our privacy protection methods to stay ahead of these changes. While the risk of re-identification using the tabulations we have published from the 2010 Census is limited, we know that the amount and accuracy of online personal information, as well the computational power to analyze that information, continues to grow.

Our recent research shows that the privacy protection methods we deployed to protect data for the 2010 Census must be improved. Our researchers have been able to simulate a "re-identification study" using publicly available data.

The accuracy of the data our researchers obtained from this study is limited, and confirmation of reidentified responses requires access to confidential internal Census Bureau information.

Nevertheless, our internal team demonstrated that an external re-identification study can match about half of the people enumerated in the 2010 Census to commercial and other online information. However, more than half of these matches are incorrect, and an external attacker has no means of confirming them. We are continuing to engage the scientific community and all stakeholders to optimize our techniques for upcoming data products including data from the 2020 Census. Our differential privacy methods will be designed to preserve the utility of our legally mandated data products while also ensuring that every respondents' personal information is fully protected.

This entry was posted on February 15, 2019 and filed under 2010 Census [/newsroom/blogs/random-samplings.html/category/Program/demosurvey/decennial/2010Census] , 2020 Census [/newsroom/blogs/randomsamplings.html/category/Program/demo-survey/decennial/2020-census] , Census Operations [/newsroom/blogs/random-samplings.html/category/Topic/censusoperations] and Disclosure Avoidance [/newsroom/blogs/randomsamplings.html/category/Topic/research/disclosure-avoidance] . Case 3:21-cv-00211-RAH-ECM-KCN Document 94-8 Filed 04/20/21 Page 1 of 4

# Exhibit 8

## 2020 Disclosure Avoidance System Updates

The Census Bureau is working closely with our data users as we modernize the privacy protections for the 2020 Census. We are reporting 2020 Disclosure Avoidance System (DAS) developments here, in our blogs [/about/policies/privacy/statistical\_safeguards/disclosure-avoidance-2020-census.html#blogs], and in our digital newsletter (Subscribe

[https://public.govdelivery.com/accounts/USCENSUS/signup/15409] | Archived Issues [/programssurveys/decennial-census/2020-census/planning-management/2020-census-dataproducts/newsletters.html] ).

We appreciate your engagement and encourage you to email comments and suggestions to 2020DAS@census.gov [mailto:2020DAS@census.gov]

EXPAND ALL | COLLAPSE ALL

### 4/7/2021: Meeting Redistricting Data Requirements: Accuracy Targets

- 2/23/2021: The Road Ahead: Upcoming Disclosure Avoidance System Milestones
- 2/3/2021: Fine-Tuning the Disclosure Avoidance System to Ensure Accuracy

### ^ 11/25/2020: Invariants Set for 2020 Census Data Products

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 noise infusion.

The first set of data products includes the P.L. 94-171 file (redistricting data), the Demographic Profiles, the Demographic and Housing Characteristics File, and the Congressional District Summary File. These products will be protected using the main TopDownAlgorithm (TDA) central to the Disclosure Avoidance System (DAS). We will employ other formal privacy methods to handle the remaining more detailed data tables separately to preserve the greatest degree of accuracy, as they pose difficult and unique privacy challenges.

Per the decision, the following statistics for this first set of products will be invariant at these levels of geography and higher:

- Total population (at the state and state-equivalents level [1])
- Total housing units (at the census block level)
- Number of group quarters facilities by type (at the census block level)

Although state-level data for American Indian and Alaska Native (AIAN) tribal areas will not be invariant, the Census Bureau will ensure that the statistics will be equally as accurate, if not more so, as those released in the most recent Privacy-Protected Microdata Files (PPMF) (v. 2020-11-16) [https://lnks.gd/l/eyJhbGciOiJIUzI1NiJ9.eyJidWxsZXRpbl9saW5rX2lkIjoxMDMsInVyaSI6ImJwMjpjbGlja yIsImJ1bGxldGluX2lkIjoiMjAyMDExMjUuMzExNDExNjEiLCJ1cmwiOiJodHRwczovL3d3dzIuY2Vuc3VzLm dvdi9wcm9ncmFtcy1zdXJ2ZXlzL2RIY2VubmlhbC8yMDIwL3Byb2dyYW0tbWFuYWdlbWVudC9kYXRhLX Byb2R1Y3QtcGxhbm5pbmcvMjAxMC1kZW1vbnN0cmF0aW9uLWRhdGEtcHJvZHVjdHMvcHBtZjIwMjAx MTE2**G24wev3ACMPEtDPEtZCFOYSHZXRWNDOCRMY0NHZAOBV9JYWFWPWDIWM**T EyNW1zZGVjczFjY2R0YXImdXRtX211ZG11bT11bWFpbCZ1dG1fc291cmNlPWdvdmRlbGl2ZXJ5In0.cLA7sES wcEcBdmnfTwy0ZOL-b8y114ZqIKel74bIgUU/s/600744772/br/90603021219-1], and as reflected in its associated Detailed Summary Metrics

[https://lnks.gd/l/eyJhbGciOiJIUzI1NiJ9.eyJidWxsZXRpbl9saW5rX2lkIjoxMDQsInVyaSI6ImJwMjpjbGlja yIsImJ1bGxldGluX2lkIjoiMjAyMDExMjUuMzExNDExNjEiLCJ1cmwiOiJodHRwczovL3d3dzIuY2Vuc3VzLm dvdi9wcm9ncmFtcy1zdXJ2ZXlzL2RlY2VubmlhbC8yMDIwL3Byb2dyYW0tbWFuYWdlbWVudC9kYXRhLX Byb2R1Y3QtcGxhbm5pbmcvMjAxMC1kZW1vbnN0cmF0aW9uLWRhdGEtcHJvZHVjdHMvcHBtZjIwMjAx MTE2LzIwMjAtMTEtMTYtZGF0YS1tZXRyaWNzLXRhYmxlcy54bHN4P3V0bV9jYW1wYWlnbj0yMDIwMT EyNW1zZGVjczFjY2R0YXImdXRtX21lZGl1bT1lbWFpbCZ1dG1fc291cmNlPWdvdmRlbGl2ZXJ5In0.e2mM5E cDMh2\_Nvp\_ZGCNkEUxSwdy64TV42bstEy-ED4/s/600744772/br/90603021219-l].

Our recent research verifies that this level of accuracy is achievable by allocating a separate, larger, dedicated privacy-loss budget for AIAN tribal area data and placing the data on the geographic hierarchy spine.

This treatment has other advantages over a state-level invariant. First, it ensures improved accuracy for AIAN tribal areas comparable to an invariant without undermining the privacy guarantee or jeopardizing system stability. Just as important, it better protects the confidentiality of tribal area residents in those states with very small tribal area populations, bringing those populations under the umbrella of guaranteed, quantifiable confidentiality protection.

The DSEP will consider and determine privacy-loss budgets for 2020 Census products at a separate meeting closer to the ultimate data release date.

This DSEP decision will be formally documented and posted on the 2020 Census Memorandum Series page

[https://lnks.gd/l/eyJhbGciOiJIUzI1NiJ9.eyJidWxsZXRpbl9saW5rX2lkIjoxMDUsInVyaSI6ImJwMjpjbGlja yIsImJ1bGxldGluX2lkIjoiMjAyMDExMjUuMzExNDExNjEiLCJ1cmwiOiJodHRwczovL3d3dy5jZW5zdXMuZ 292L3Byb2dyYW1zLXN1cnZleXMvZGVjZW5uaWFsLWNlbnN1cy8yMDIwLWNlbnN1cy9wbGFubmluZy1tY W5hZ2VtZW50L21lbW8tc2VyaWVzLmh0bWw\_dXRtX2NhbXBhaWduPTIwMjAxMTI1bXNkZWNzMWNjZ HRhciZ1dG1fbWVkaXVtPWVtYWlsJnV0bV9zb3VyY2U9Z292ZGVsaXZlcnkifQ.1o5zqNKmQ1IU6zJ9P8tZp wzsWLUj3H37R1RhnyL6dnc/s/600744772/br/90603021219-l].

[1] The District of Columbia and Puerto Rico

/	11/16/20: N	low Online:	Revised	<b>Privacy-Prote</b>	cted Microdata Files
---	-------------	-------------	---------	----------------------	----------------------

\* 10/23/20: Issue Discovered in the Privacy-Protected Microdata File

- ✓ 9/24/2020: New Metrics and Demonstration Data
- 9/17/20: New Privacy-Protected Microdata Files (PPMFs)
- 8/26/20: New Frequently Asked Question re: DAS Development Schedule
- **7/14/20: New Privacy-Protected Census Demonstration Data**
- 7/1/20: Census Bureau Partners with Committee on National Statistics to Produce New Demonstration Data Files

- 6/26/20:2 New Frequencies to Still of 4
- 6/1/20: New Metrics and DAS Updates Presentations from CNSTAT Expert Meeting on Disclosure Avoidance
- 5/27/20: Release of "2010 Demonstration Metrics 2;" First Set of Post-Baseline Quality Metrics Results
- 5/18/20: New 2020 Census Data Products Newsletter
- 4/20/20: New Frequently Asked Questions
- **4/17/20: Video: DAS Development Status Update**

### ✓ 3/27/20: Draft Quality Measures and Benchmark

#### **Related Information**

Developing the DAS: Demonstration Data and Progress Metrics [/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das development.html]

2020 Disclosure Avoidance System: Frequently Asked Questions [/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-daupdates/2020-das-faqs.html]

Disclosure Avoidance and the 2020 Census [/about/policies/privacy/statistical\_safeguards/disclosure-avoidance-2020-census.html]

Last Revised: April 8, 2021

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-9 Filed 04/20/21 Page 1 of 5

## Exhibit 9

### UNITED STATES DISTRICT COURT FOR THE MIDDLE DISTRICT OF ALABAMA

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,

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,

٧.

No. 3:21-CV-211-RAH-ECM-KCN

Defendants.

### SUPPLEMENTAL DECLARATION OF CLAY S. HELMS

My name is Clay S. Helms. I am the Deputy Chief of Staff and Director of Elections for the Alabama Secretary of State's Office. I am over the age of 19 and the facts I have set out below are based upon my personal knowledge or the records of the Alabama Secretary of State's Office maintained in the ordinary course of business. I have read and reviewed Defendants' Response, Doc. 41, and the Declaration of James Whitehorne, Doc. 41-3. I am providing this declaration as a supplement my earlier declaration, *see* Doc. 3-3, and to correct misconceptions in the Census Bureau's Response, Doc. 41, and in Whitehorne's Declaration, Doc. 41-3.

In my earlier declaration, I stated that the delay caused by the Bureau's February 12
Decision "could result in one or more 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." Doc.  $3-3 \ 12$ .

- 2. By using the word "could," I meant to suggest that the result *could* be at least one of those three eventualities. I did not mean to suggest these eventualities are not likely, *contra* Doc. 41 at 37. Rather, I believe that at least one of these eventualities is very likely to occur—and that any combination of them could occur—if the Bureau forces the Boards of Registrars to reassign voters on the compressed timeline the Bureau has suggested. A rushed reassignment process is very likely to lead to mistaken reassignments, for example, considering that a significant percentage of Registrars have never participated in a decennial voter-reassignment process before. Additionally, it is substantially likely that if the Bureau delays delivery of redistricting data as it has promised to do—thus delaying the State's redistricting process and thus the Boards of Registrars' ability to reassign voters—the compressed schedule will force at least some Boards of Registrars to enter into expensive contracts with outside vendors to assist in the voter-reassignment process.
- 3. The Bureau also asserts that its delay in delivering redistricting data does not harm the Boards of Registrars by reducing their time to reassign voters because "the Boards of Registrars can get started now with information that the Census Bureau has already provided." Doc. 41 at 37. Specifically, the Bureau suggests that States can use specific data provided to "prepare voter rolls." Whitehorn Decl., Doc. 41-3 ¶ 12.
- 4. I disagree with the Bureau's statement. To the best of my knowledge, the Boards of Registrars' practice has always been to use decennial census redistricting data to reassign voters. Many of the Boards of Registrars therefore may not be able to reassign voters or

prepare voter rolls without redistricting data—and, to the extent the Boards are able to reassign voters without the data, they will have a much more difficult time doing so. Following the Bureau's suggestion would therefore be substantially likely to increase the risk of erroneous voter reassignment. It is my view that such risks in turn jeopardize the State's election integrity and the public's confidence.

- 5. Additionally, "get[ting] started" on voter reassignment now, without the Bureau's redistricting data, is likely to produce inaccurate and ultimately unusable reassignments. Assigning voters to census blocks now, then later assigning voters to districts after the Legislature draws new districting plans, would place people who moved between those dates into incorrect districts.
- 6. So to the extent that the Boards could perform any part of the process now, the changes in precinct and district boundaries following redistricting are substantially likely to require the Boards to redo parts of the process again later to ensure their continued accuracy, which would require incurring duplicative costs and man-hours.
- 7. Lastly, the Bureau stated that I "ma[de] clear that the State can 'push[] back [its] primary election' by seven weeks." Doc. 41 at 37. The Bureau's selective edits materially misstated my declaration to suggest that this is an option I endorse. I dispute this characterization. I explained in my previous declaration that "Alabama may have to reschedule the primary and runoff elections," and cautioned that while this strategy "may mitigate the previously discussed harms, it does not eliminate them. A seven-week delay by the State cannot entirely offset the harms caused by the Census Bureau's six-month delay in providing redistricting data." Doc. 3-3 ¶¶ 14-15 (emphasis added). I still believe Alabama may have no choice but to reschedule its elections, but delaying elections would involve its own

n

hassles, costs, and difficulties, and still would not completely remedy the injuries caused by the Bureau's delay. These are issues that the Legislature would have to weigh in deciding whether to exercise that option.

Pursuant to 28 U.S.C. § 1746, I declare under penalty of perjury that the foregoing is true and correct. Executed on April 16, 2021.

- lay 5.

Clay S. Helms Deputy Chief of Staff & Director of Elections Office of the Secretary of State State of Alabama

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-10 Filed 04/20/21 Page 1 of 3

# Exhibit 10

### John M. Abowd

Home

**Professional Information** 

**Recent News** 

Special Materials

## Tweetorial: Reconstruction-abetted reidentification attacks and other traditional vulnerabilities

Courses

- 1. First, let's get the facts straight: the U.S. Census Bureau reconstructed 100% of the 2010 Census micro-data records (308,745,538 persons).
- 2. Those records contained: full census block id (all 15 digits), voting age (yes/no), sex, age (in years), race (all 63 OMB categories), and ethnicity (Hispanic or not).
- 3. The reconstructed records matched the confidential data (2010 CEF) exactly (every single bit) for 46% of the population (142 million people) and allowing age +/- 1 year for 71% of the population (219 million people).
- 4. Those match rates are salient because in the confidential data, more than 50% of the population is unique on those variables (block, sex, age, race and ethnicity)
- 5. This makes the confidential data vulnerable to a re-identification attack by linkage to external data with some or all of the same variables.
- 6. This is precisely the reason that the Census Bureau has never released public-use micro-data with detailed geography, but if you can reconstruct the detailed geography from the published tables, the vulnerability is still there.
- 7. Using commercial databases harvested between 2009 and 2011 in support of the 2010 Census, the Census Bureau linked PII (name and address-technically PIK and MAFID) to the reconstructed micro-data.
- 8. This linkage resulted in putative re-identification of 138 million persons (45% of the population).
- 9. This estimate of the success rate (also called the recall rate) is almost certainly conservative because neither the 2010 Census nor the commercial databases has PII for all 309 million persons.
- When the Census Bureau linked the PII-laden reconstructed data (putative re-identifications) to the 2010 Census CEF, it confirmed the correctness of 52 million persons (confirmation rate 38%).
- 11. This confirmation rate is also conservative because no use was made of the relationship-tohouseholder or household composition data in the published tables.
- 12. The last time Census Bureau researchers published a re-identification study, the putative reidentification rate was 0.017% (389 of 2.3 million), and the confirmation rate was 22% (87 of 389).
- 13. That's an aggregate vulnerability (product of the two rates) of 0.0038%.
- 14. That aggregate vulnerability for the 2010 Census, based on these more recent studies, turns out to be 17%-four orders of magnitude greater.
- 15. Re-identification risk is only one part of the Census Bureau's statutory obligation to protect confidentiality. The statute also requires protection against exact attribute disclosure.
- 16. Neither the census block nor voting age received any confidentiality protection in the tabular summaries from the 2010 Census (this is public information in all of the relevant technical documentation).
- 17. Consequently, the micro-data reconstruction of block and voting-age is always exactly correct, and exactly matches the confidential data.

<b>^</b> 1	
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	

### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-10 Filed 04/20/21 Page 3 of 3

- 18. Any block where the voting-age data are either all "yes" or all "no" is an exact attribute disclosure assignable to all persons living in that block on April 1, 2010.
- 19. There is wide-spread recognition in the official statistics community that both reconstructionabetted re-identification and reconstruction-abetted exact attribute disclosure are unacceptable vulnerabilities for 2020 Census publications,
- 20. Those publications may include a block-level citizen voting age population by race and ethnicity table.
- 21. Former Census Bureau Director John Thompson and former BLS Commissioner Erica Groshen have both publicly said that these vulnerabilities must be addressed.
- 22. Differential privacy, as implemented for the 2020 Census, directly addresses both of these traditional vulnerabilities, and allows the publisher to manage the accuracy of the resulting tables to ensure fitness-for-use.
- 23. No traditional SDL method can make that claim, accompanied by proof.
- 24. You are free to take issue with this risk assessment, but the statutory confidentiality protection obligation is the domain of the Census Bureau, and the protections of Title 13, section 9 are not subject to a "when convenient" exception.
- 25. This tweetorial and slides from my recent AAAS and AAG talks published here.

Hosted by CampusPress

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-11 Filed 04/20/21 Page 1 of 6

# Exhibit 11

To: Nathaniel Cogley (CENSUS/DEPDIR EED)[nathaniel cogley@census.gov]\_ile\_104/20/21, Benjamin A Ovemoit (CENSUS/DEPDIR FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=6681E88243DE4600871932859CC06C9F-OVERHOLT, B] Sent: Thur 10/22/2020 7:11:24 PM (UTC)

I have two concerns with the misstatements and inappropriate promises offered in the third Paragraph. First, we state that we have already decided to change the DAS geographic Hierarchy in favor of AIAN groups. Second, it suggests we are seriously considering allocating more of the privacy loss budget to the AIAN community. These statements are either false or need to be reconsidered.

To my first concern, as I understand it, the Director and Secretary are still deciding how, where, and even if we will employ Disclosure Avoidance. Stating that "We are making changes to the 2020 DAS geographic hierarchy to more effectively ensure accuracy for AIAN tribal areas." before we have made the decisions on DA to begin with, is misleading at best. At worst, it invites future lawsuits from any other minority group that is disadvantaged relative to the AIAN community.

To my second concern, it is important to understand the extent to which minority groups, state, and even the statistical community, opposes the use of Disclosure avoidance. By trying to placate one racial group with a promise of "working to improve the accuracy of population counts in AIAN tribal areas by allocating more of the privacy-loss budget to those statistics." at the cost of less accuracy in the population counts of all other minority groups in all other areas, is an invitation extensive legal action.

Together, the issues I have presented amount to Associate Directors making decisions that are the sole prerogative of the Director and the Secretary. We cannot promise to do something that blatantly gives one racial group an advantage at the expense of all others. Right now, the AIAN community is engaged against our proposed methodology, this letter guarantees every other minority group will become more engaged and demand equal treatment.

-Ben

From: Nathaniel Cogley (CENSUS/DEPDIR FED) <nathaniel.cogley@census.gov> Sent: Thursday, October 22, 2020 2:35 PM To: Benjamin A Overholt (CENSUS/DEPDIR FED) <benjamin.a.overholt@census.gov> Subject: Fw: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

From: Steven K Smith (CENSUS/DEPDIR FED) <steven.k.smith@census.gov>

Sent: Thursday, October 22, 2020 1:30 PM

**To:** Nathaniel Cogley (CENSUS/DEPDIR FED) <nathaniel.cogley@census.gov>; Earl N Mayfield (CENSUS/DEPDIR FED) <earl.n.mayfield@census.gov>

**Subject:** Fw: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

Any comments?

From: Christopher J Stanley (CENSUS/OCIA FED) <christopher.j.stanley@census.gov>

Sent: Thursday, October 22, 2020 1:24 PM

To: Steven K Smith (CENSUS/DEPDIR FED) < steven.k.smith@census.gov>; Ali Mohammad Ahmad (CENSUS/ADCOM FED)

<ali.m.ahmad@census.gov>; Michael John Sprung (CENSUS/DEPDIR FED) <michael.j.sprung@census.gov>

**Subject:** Re: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

From: Christopher J Stanley (CENSUS/OCIA FED) <christopher.j.stanley@census.gov>
Sent: Friday, October 9, 2020 11:02 AM
To: Steven K Smith (CENSUS/DEPDIR FED) <steven.k.smith@census.gov>; Ali Mohammad Ahmad (CENSUS/ADCOM FED)
<ali.m.ahmad@census.gov>
Subject: Re: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

The letter to Haaland and Cole might require more edits than I initially thought.

I don't like "we can't postpone decisions" because we have done that twice now, at least in regards to the DSEP meeting. So I rephrased.

For the tribal consultation, the draft Federal Notice for that is now with Caryn, so Smith, you might already have it. If you are able to expedite that, it could me published probably next week.

Maybe we hold this letter just a little bit longer. It could include more specifics on the tribal consultation if the FRN is approved and then sent along for publication. Edits that could be edited again if the FRN can be finalized are attached.

From: Christopher J Stanley (CENSUS/OCIA FED) <christopher.j.stanley@census.gov>

Sent: Friday, October 9, 2020 9:38 AM

**To:** Steven K Smith (CENSUS/DEPDIR FED) <steven.k.smith@census.gov>; Ali Mohammad Ahmad (CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>

**Subject:** Re: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

For Warren, Sprung told me that yesterday and I let the correspondence people know. thank you.

I might need minor edits to Haaland / Cole letter. I am working it right now and will send with track changes if anything needs to be edited. I think the disclosure avoidance stuff is probably still good to go, but the tribal consultation part might need some refinements now.

From: Steven K Smith (CENSUS/DEPDIR FED) <steven.k.smith@census.gov>
Sent: Friday, October 9, 2020 9:34 AM
To: Ali Mohammad Ahmad (CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>
Cc: Christopher J Stanley (CENSUS/OCIA FED) <christopher.j.stanley@census.gov>
Subject: Re: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

Steve approved letter to Sen Warren yesterday...

From: Ali Mohammad Ahmad (CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>
Sent: Friday, October 9, 2020 9:25 AM
To: Steven K Smith (CENSUS/DEPDIR FED) <steven.k.smith@census.gov>
Subject: Re: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

It looks accurate but Stanley is reviewing one more time to makes sure it's the latest version.

Ali Ahmad, Associate Director Communications Directorate U.S. Census Bureau O: 301-763-8789 | M: 240-532-0676 Ali.M.Ahmad@census.gov census.gov | @uscensusbureau

From: Steven K Smith (CENSUS/DEPDIR FED) <steven.k.smith@census.gov>
Sent: Friday, October 9, 2020 9:15 AM
To: Ali Mohammad Ahmad (CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>
Subject: Fw: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

Ali:

Are the dates and milestones reference in this letter still appropriate? Thanks.

From: Caryn M Tate (CENSUS/DEPDIR FED) <caryn.m.tate@census.gov>
Sent: Friday, October 9, 2020 9:05 AM
To: Steven K Smith (CENSUS/DEPDIR FED) <steven.k.smith@census.gov>
Subject: Fw: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

Good morning Steve,

Could you help me get this cleared by the Director?

Thanks, Caryn

**Caryn Tate,** Office Manager, Office of the Director, U. S. Census Bureau Office: 301-763-1138 Fax: 301-763-3761 caryn.m.tate@census.gov census.gov Connect with us on Social Media

From: Michael John Sprung (CENSUS/DEPDIR FED) <michael.j.sprung@census.gov> Sent: Thursday, September 10, 2020 1:47 PM To: Caryn M Tate (CENSUS/DEPDIR FED) <caryn.m.tate@census.gov>; Katherine Dodson Hancher (CENSUS/DEPDIR FED) <Katherine.Dodson.Hancher@census.gov> Subject: Fw: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

This looks fine to me. Approved.

From: Katherine Dodson Hancher (CENSUS/DEPDIR FED) <Katherine.Dodson.Hancher@census.gov>
Sent: Wednesday, September 9, 2020 2:26 PM
To: Michael John Sprung (CENSUS/DEPDIR FED) <michael.j.sprung@census.gov>
Cc: Caryn M Tate (CENSUS/DEPDIR FED) <caryn.m.tate@census.gov>

Subject: Fw: FOR THE DEPLITY DIRECTOR/DIRECTOR'S APPROVAL: COAS-10523 Deb Haaland (D-MM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and Al/AN Data from the Native American Caucus.

Mike-

For your review and approval. Please respond to Caryn and myself with attachments.

Kathy

Kathy Hancher Office of the Director U.S. Census Bureau 301.763.3964 katherine.dodson.hancher@census.gov

From: Ron S Jarmin (CENSUS/DEPDIR FED) <Ron.S.Jarmin@census.gov>
Sent: Tuesday, September 8, 2020 5:20 PM
To: Caryn M Tate (CENSUS/DEPDIR FED) <caryn.m.tate@census.gov>; Katherine Dodson Hancher (CENSUS/DEPDIR FED)
<Katherine.Dodson.Hancher@census.gov>
Subject: Fw: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

Approved. Signed control sheet attached.

Ron S Jarmin, PhD., Deputy Director U.S. Census Bureau o: 301-763-1858 | m: 301-980-8140 census.gov | @uscensusbureau Shape your future. START HERE > 2020census.gov

From: Katherine Dodson Hancher (CENSUS/DEPDIR FED) <Katherine.Dodson.Hancher@census.gov>
Sent: Tuesday, September 8, 2020 4:27 PM
To: Ron S Jarmin (CENSUS/DEPDIR FED) <Ron.S.Jarmin@census.gov>
Cc: Caryn M Tate (CENSUS/DEPDIR FED) <caryn.m.tate@census.gov>
Subject: Fw: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

For your approval. Please respond to Caryn and myself with attachments.

Kathy Hancher Office of the Director U.S. Census Bureau 301.763.3964 katherine.dodson.hancher@census.gov

From: BOC Correspondence Quality Assurance (CENSUS) <boc.correspondence.quality.assurance@census.gov><br/>Sent: Tuesday, September 8, 2020 4:26 PM

**To:** Caryn M Tate (CENSUS/DEPDIR FED) <caryn.m.tate@census.gov>; Katherine Dodson Hancher (CENSUS/DEPDIR FED) <Katherine.Dodson.Hancher@census.gov>; Josie A Hollingsworth (CENSUS/DEPDIR FED) <josie.a.hollingsworth@census.gov>; Natalie M Jackson (CENSUS/EMD FED) <Natalie.M.Jackson@census.gov>

Cc: BOC Correspondence Quality Assurance (CENSUS) < boc.correspondence.quality.assurance@census.gov>; Christopher J Stanley
(CENSUS/OCIA FED) < christopher i stanley@census.gov?CN Document 94-11 Filed 04/20/21 Page 6 of 6 Subject: FOR THE DEPUTY DIRECTOR/DIRECTOR'S APPROVAL: CQAS-10523 Deb Haaland (D-NM-01) Tom Cole (R-OK-04) Letter regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus.

Hi all,

The attached draft to Congressional members Haaland/Cole regarding Census Disclosure Avoidance System and AI/AN Data from the Native American Caucus is ready for the Deputy Director/ Director's approval. Chris, Ali, and Christa have all approved. It is reflected in the control sheet. Please let us if they concur.

Thanks. Nicole Case 3:21-cv-00211-RAH-ECM-KCN Document 94-12 Filed 04/20/21 Page 1 of 13

# Exhibit 12

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

THE STATE OF ALABAMA, et al.

Plaintiffs,

v.

No. 3:21-CV-00211- RAH-ECM-KCN

UNITED STATES DEPARTMENT OF COMMERCE, *et al.*,

Defendants.

# DEFENDANTS' RESPONSES TO PLAINTIFFS' FIRST REQUEST FOR ADMISSIONS

Defendants United States Department of Commerce; Gina Raimondo, in her official capacity as Secretary of Commerce; United States Census Bureau; and Ron Jarmin, in his official capacity as Acting Director of the U.S. Census Bureau; collectively submit the following responses and objections to Plaintiffs' First Set of Requests for Admissions. Responses to the following Requests for Admissions are subject to the general objections set forth at the end of this document.

**REQUEST FOR ADMISSION NO. 1**: Admit that Differential Privacy has never been used as the disclosure avoidance method in any decennial census before the 2020 census.

<u>ANSWER</u>: ADMIT, although Defendants note that the Census has used differential privacy to protect confidential information in other Census programs.

**REQUEST FOR ADMISSION NO. 2**: Admit that the implementation of Differential Privacy has, at least in part, delayed the release of the PL 94-171 Data to Alabama. **ANSWER**: DENY. Defendants specifically DENY that applying any other appropriate disclosure-avoidance methods that complied with Title XIII would allow them to release Redistricting Data any sooner than using Differential Privacy. In fact, switching to a disclosure avoidance methodology other than Differential Privacy would require new disclosure risk analysis, software development, system testing, and quality assurance steps that would result in substantial delays to the current production schedule for the Redistricting Data.

**REQUEST FOR ADMISSION NO. 3**: Admit that after the Census Bureau performed an internal database reconstruction and re-identification attack on the 2010 census data, John M. Abowd—the Chief Scientist and Associate Director for Research and Methodology at the U.S. Census Bureau—concluded that "the risk of re-identification" posed by such an attack "is small."

<u>ANSWER</u>: Defendants object to the Request's selective quote as lacking context and as written, potentially misleading because Dr. John M. Abowd specifically retracted this claim in February 16, 2019 in his presentation to the American Association for the Ad-vancement of Science because subsequent research revealed that 52 million persons, at a minimum, could be correctly re-identified from the reconstructed 2010 Census micro-data.

ADMIT only to the extent that Dr. John M. Abowd, made a presentation on August 23, 2018, to the 24<sup>th</sup> ACM SIGKDD Conference on Knowedge Discovery and Data Mining, titled "The U.S. Census Bureau Adopts Differential Privacy" where he: described the Census Bureau's internal experiments that "[c]onfirm[ed] that the confidential microdata from the confidential hundred percent file can be reconstructed quite accurately from PL94 + balance of SF1"; stated that "[w]hile there is a vulnerability, the risk of reidentification is small"; and concluded that "reconstruction of Title 13-sensitive data is an issue," thus providing "[s]trong motivation for the adoption of differential privacy for the 2018 End-to-End Census Test and 2020 Census." The context of the quote can be seen in the slide below.



Otherwise, DENY.

**REQUEST FOR ADMISSION NO. 4**: Admit that to the extent the Census Bureau's internal database reconstruction and re-identification attack on the 2010 census data revealed some risk of reidentification, the Bureau's application of Differential Privacy to the 2010 census data did not remove all risk of reidentification.

**ANSWER**: ADMIT only to the extent that the Census Bureau's internal database reconstruction and re-identification attack on the 2010 census data revealed that a hypothical attacker could correctly identify at least 52 million Americans and potentially up to 179 million. And ADMIT only to the extent that releasing any data, regardless of the disclosure avoidance method employed, involves some risk of reidentification. The Census conducted an empirical analysis that showed differential privacy offered the most

efficient trade-off between privacy and accuracy and its calculations showed that the efficiency of differential privacy dominated traditional methods. 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 Bureau, including suppression and data swapping. The Census Bureau has assessed various options for mitigating the risk of reidentification in Redistricting Data and concluded that Differential Privacy strikes the most appropriate balance between privacy, end-user needs, and accuracy. Otherwise, DENY.

**REQUEST FOR ADMISSION NO. 5**: Admit that with Differential Privacy applied, the data at the census block level will be less reflective of both the number of persons and the demographic characteristics of persons in that census block than without Differential Privacy applied.

**ANSWER**: Defendants object to this Request as vague and ambiguous. The phrase "less reflective" is vague and imprecise, particularly when coupled with the compound phrase "both the number of persons and demographic characteristics." Defendants object to this Request's phrase "without Differential Privacy applied" as vague and ambiguous because it is unclear whether the phrase refers to collected and imputed demographic data without any disclosure avoidance applied, or collected and imputed demographic data with some other disclosure avoidance applied, or the actual population count and demographics that existed on Census Day.

LACK OF INFORMATION. Defendants can neither admit nor deny this Request because the Census has not finalized the specific algorithm parameters and other critical details about how it will apply Differential Privacy to the Redistricting Data. Further, the enumerated population count and collected demographic characteristics for any given census block are imperfect representations of the actual population count and demographic characteristics of that census block that existed on Census Day. So for certain census blocks, application of Differential Privacy may result in a population count and demographic characteristics that are just as reflective – or in some cases even more reflective – "of both the number of persons and the demographic characteristics of persons in that census block" as the confidential census data themselves.

**REQUEST FOR ADMISSION NO. 6**: Admit that there is not a single documented case of anyone outside the Census Bureau obtaining the responses of a particular identified person in the public use decennial census or American Community Survey data.

<u>ANSWER</u>: DENY. See the documented cases in McKenna (2019) <u>Research and</u> <u>Methodology Directorate (census.gov)</u>.

**REQUEST FOR ADMISSION NO. 7**: Admit that if 13 U.S.C. § 9 requires the Bureau to protect only respondents' identities, and not also their characteristics, then Differential Privacy would not be required to satisfy 13 U.S.C. § 9's confidentiality requirement.

ANSWER: Defendants object to this Request as an improper hypothetical request for admission. *See, e.g., Atlanta Channel, Inc. v. Solomon, Civil Action,* No. 15-1823 (RC), 2020 U.S. Dist. LEXIS 216969, at \*18 (D.D.C. Nov. 18, 2020); *Storck USA, L.P. v. Farley Candy Co.,* No. 92 C 552, 1995 U.S. Dist. LEXIS 4433, at \*9 (N.D. Ill. Apr. 5, 1995); *Morley v. Square, Inc.,* No. 4:14cv172, 2016 U.S. Dist. LEXIS 318, at \*11 (E.D. Mo. Jan. 4, 2016); *Buchanan v. Chi. Transit Auth.,* No. 16-cv-4577, 2016 U.S. Dist. LEXIS 168983, at \*15 (N.D. Ill. Dec. 7, 2016).

Defendants CANNOT ADMIT OR DENY this Request because it is premised on a mischaracterization of the law. 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).

AMENDED REQUEST FOR ADMISSION NO. 8: Admit that the statements made at the following governmental web address are authentic and accurately reflect the Bureau's position as to the implementation of Differential Privacy with respect to the Census: <u>https://ask.census.gov/prweb/PRServletCustom/app/ECOR-</u>

RAsk/YACFBFye-

<u>rFIz\_FoGtyvDRUGg1Uzu5Mn\*/!STANDARD?pzuiactionzzz=CXtpbn0rTEpMcGRYOG</u> <u>1vS0tqTFAwaENUZWpvM1NNWEMzZ3p5aFpnWUxzVmw0TjJp-</u>

TUZqanNtWnVVTitZWnhVVldtN1dI\*.

ANSWER: ADMIT, presuming that the Request refers to text in the following im-

age.

# Has the Census Bureau settled on final policy for differential privacy?

Yes, the Census Bureau will protect 2020 Census data using the cryptography-based disclosure avoidance system known as differential privacy.

The Census Bureau has a dual mandate to produce quality statistical information while protecting the confidentiality of respondent data. We know that the nation needs timely and accurate information to make informed decisions. Our goal is to ensure that the public trusts us with their data, and values the statistics that we produce. Adopting our advanced confidentiality protection system helps us to meet that goal.

See also: Disclosure Avoidance and the 2020 Census

SHARE ARTICLE

**REQUEST FOR ADMISSION NO. 9**: Admit that with Differential Privacy applied, areas with comparatively smaller populations—such as rural areas—will see a greater variance in population figures from the raw data than will areas with comparatively larger populations—such as urban areas.

ANSWER: LACK OF INFORMATION. Defendants can neither admit nor deny

this Request because the Census has not finalized the specific algorithm parameters and other critical details about how it will apply Differential Privacy to the Redistricting Data.

**REQUEST FOR ADMISSION NO. 10**: Admit that with Differential Privacy applied, smaller subpopulations, such as specific racial groups, will see a greater variance from the raw data than more numerous racial or ethnic groups.

<u>ANSWER</u>: LACK OF INFORMATION. Defendants can neither admit nor deny this Request because the Census has not finalized the specific algorithm parameters and other critical details about how it will apply Differential Privacy to the Redistricting Data.

**REQUEST FOR ADMISSION NO. 11**: Admit that Differential Privacy adds intentional error to the PL 94-171 Data.

<u>ANSWER</u>: ADMIT that most disclosure avoidance methods, including the swapping methodology used for the 2000 and 2010 Censuses and differential privacy, add intentional error to the PL 94-171 data.

**REQUEST FOR ADMISSION NO. 12**: Admit that demographic information (such as age, race, ethnicity, etc.) will not be held invariant at any level of geography under Differential Privacy for the 2020 Census.

<u>ANSWER</u>: LACK OF INFORMATION. Defendants can neither admit nor deny this Request because the Census has not finalized the specific algorithm and other critical details about how it will apply Differential Privacy to the Redistricting Data. While the Census Bureau's Data Stewarship Executive Policy (DSEP) Committee announced a final list of data elements that will be held invariant for the 2020 Census in November 2020, those invariants are always subject to change up to the point when the Secretary of Commerce releases the Redistricting Data.

7

**REQUEST FOR ADMISSION NO. 13**: Admit that with Differential Privacy applied, the PL 94-171 data will not reflect the actual population counts for sub-statewide geographies.

<u>ANSWER</u>: Defendants object to this Request as vague and ambiguous because "actual population counts" could refer to the actual population that existed on Census Day, or different types of collected information, including self-response, persons enumerated in non-response follow-up, proxy responses, group quarters response, or imputed population counts, or some subset of those categories.

LACK OF INFORMATION. Defendants can neither admit nor deny this Request because the Census has not finalized the specific algorithm parameters and other critical details about how it will apply Differential Privacy to the Redistricting Data.

**REQUEST FOR ADMISSION NO. 14**: Admit that if the Census Bureau did not apply Differential Privacy, the Bureau could shorten the timeline for the release of redistricting data.

#### ANSWER: DENY.

**REQUEST FOR ADMISSION NO. 15**: Admit that if the Census Bureau did not apply Differential Privacy, the Bureau could release the PL 94-171 data to states on a stateby-state rolling basis, rather than in a single national release.

ANSWER: Defendants object to this Request as an improper hypothetical request for admission. *See, e.g., Atlanta Channel, Inc. v. Solomon, Civil Action,* No. 15-1823 (RC), 2020 U.S. Dist. LEXIS 216969, at \*18 (D.D.C. Nov. 18, 2020); *Storck USA, L.P. v. Farley Candy Co.,* No. 92 C 552, 1995 U.S. Dist. LEXIS 4433, at \*9 (N.D. Ill. Apr. 5, 1995); *Morley v. Square, Inc.,* No. 4:14cv172, 2016 U.S. Dist. LEXIS 318, at \*11 (E.D. Mo. Jan. 4, 2016); *Buchanan v. Chi. Transit Auth.,* No. 16-cv-4577, 2016 U.S. Dist. LEXIS 168983, at \*15 (N.D. Ill. Dec. 7, 2016).

8

To the extent that the Request asks Defendants to admit that applying Differential Privacy prevents the Bureau from releasing PL 94-171 data on a rolling basis, Defendants DENY. The Census Bureau can release data on either on a state-by-state rolling basis or a single, national release regardless of what disclosure avoidance method it employs. Census Bureau could release PL 94-171 data to states on a state-by-state rolling basis after applying Differential Privacy rather than a single national release, but as explained in the February 12 Press Release, a single national release of final redistricting "enables the Census Bureau to deliver complete and accurate redistricting data in a more timely fashion overall for the states." *See* Pls. Mot. for P.I., Ex. 7. To the extent the Census Bureau does not apply Differential Privacy, it would need to apply an alternate form of disclosure avoidance that complies with Title XIII. But whatever hypothetical Title XIII-compliant disclosure-avoidance mechanism the Census Bureau uses, the form of disclosure avoidance mechanism would most likely not impact whether data can be released on a rolling basis or on a single national release.

**REQUEST FOR ADMISSION NO. 16**: Admit that Differential Privacy has never before, to the Census Bureau's knowledge, been applied to a dataset as large and complex as the Census.

<u>ANSWER</u>: DENY. Differential Privacy is used by companies that manage extremely large data-sets, including Google, Apple, and Facebook. Because those private data-sets collect more information than the Census, they are likely larger and more complex than the redistricting data set collected by the Census.

# **GENERAL OBJECTIONS**

Defendants object to the definition of "you" and "your," and Instruction
 No. 5, to the extent they include persons and entities that are not parties to this lawsuit.

9

#### Case 3:21-cv-00211-RAH-ECM-KCN Document 94-12 Filed 04/20/21 Page 11 of 13

Defendants construe "you" and "your" to refer to the persons and entities named as defendants in this case.

2. Defendants object to the Definitions and Instructions to the extent they purport to impose obligations on Defendants that go beyond the obligations imposed by those the Federal Rules of Civil Procedure. Pursuant to Fed. R. Civ. P. 36(a)(3) and 29, the parties have agreed that Defendants will provide responses to these Requests for Admission by April 13, 2021.

3. Defendants object to Instruction No. 2 to the extent it purports to impose obligations on Defendants that go beyond the obligations those in Fed. R. Civ. P. 26(e). Defendants will timely supplement their responses as necessary and consistent with their obligations under Fed. R. Civ. P. 26(e). DATED: April 13, 2021

BRIAN M. BOYNTON Acting Assistant Attorney General

ALEXANDER K. HAAS Director, Federal Programs Branch

BRAD P. ROSENBERG Assistant Director, Federal Programs Branch

<u>/s/Zachary A. Avallone</u> ZACHARY A. AVALLONE ELLIOTT M. DAVIS JOHN ROBINSON Trial Attorneys Civil Division, Federal Programs Branch U.S. Department of Justice 1100 L St. NW Washington, DC 20005 Phone: (202) 514-2705 Fax: (202) 616-8470 E-mail: zachary.a.avallone@usdoj.gov

Counsel for Defendants

# CERTIFICATE OF SERVICE

I hereby certify that on April 13, 2021, I served Defendants' Responses to Plaintiffs'

First Request for Admissions to counsel of record for the Plaintiffs via e-mail per

agreement.

DATED: April 13, 2021

<u>/s/ Elliott M. Davis</u> 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

Counsel for Defendants

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-13 Filed 04/20/21 Page 1 of 16

# Exhibit 13

# Census Scientific Advisory Committee Fall 2020 Meeting Recommendations

To:	Ron S. Jarmin
	Acting Director
	U.S. Census Bureau
From:	Allison Plyer
	Chair
	Census Scientific Advisory Committee
Subject:	Partial Recommendations and Comments to the U.S. Census Bureau from the Census Scientific Advisory Committee Fall 2020 Meeting
September	17-18, 2020

The Census Scientific Advisory Committee (CSAC) thanks the U.S. Census Bureau for their thorough planning and preparation for this first ever virtual CSAC meeting. The topics covered were timely and salient. The presenters were enthusiastic and engaging. And with very few glitches the technology worked well to support discussion among Census staff and CSAC members participating from remote locations around the country. While in-person meetings support greater information exchange and optimal communication, the virtual platform worked will give the need for physical distancing under pandemic conditions. CSAC thanks the Bureau staff for their extraordinary efforts to make this meeting possible.

# Update on the 2020 Census Operations

The Census Scientific Advisory Committee (CSAC) commends the Census Bureau on executing the operations of the 2020 Census in the midst of an unexpected global pandemic. Field work timelines were rewritten from scratch. Contact attempts were redesigned in a short time so that field work was safe for both Bureau employees as well as for respondents. Even the communications campaign was adapted, and new advertising generated in some instances. And the Bureau nimbly executed many advances to plan accelerated post data collection processes. Throughout this process, CSAC has been impressed by the dedication of Census Bureau staff.

After almost a decade of planning, the pandemic outbreak occurred just as field work was starting. In this context, the Census Bureau was able to quickly adapt to changing circumstances and execute data collection in a way that is largely consistent with its planned operational goals. For example, the Bureau was able to meet its goal of having 60.5% of housing units self-respond through mail (as had been done in previous censuses), telephone or internet (two new modes of data collection). Internet self-response was critical to achieving

this goal, particularly given the dynamic of the pandemic. We applaud the Census Bureau on the success of the use of this new technology.

Census methodology requires the Bureau now resolve all remaining housing units through ongoing Non-Response Follow-Up (NRFU), and then complete a series of complex post data collection processes, some of which remain untested and untried.

1) To ensure a successful completion of the 2020 Census in a way that is consistent with its mandate of counting everyone once and in the right place, and based on its scientific and methodological expertise, CSAC recommends that the 2020 Census operational timeline be extended per the Bureau's April 2020 request. Counting everyone once and in the right place, using untested and never-before-used technologies, that must work together with precision, requires time. When the weather isn't right, we postpone the launching of rockets into space. The same should be true of the decennial enumeration, the results of which will impact apportionment, redistricting, funding decisions, legal mandates, and regulatory uses of decennial Census data over the next decade.

# **CENSUS BUREAU RESPONSE:**

# *The Census Bureau does not accept this recommendation for the 2020 Census. Data collection was completed on October 15, 2020.*

- 2) Based on discussions during the Fall 2020 CSAC meeting, the risks to data accuracy from a compressed timeline are substantial. CSAC advises the Census Bureau that the following issues may compromise the accuracy of the 2020 Census under the "Replan" compressed timeline for Non-Response Follow-up:
  - On-going events, including natural disasters and civil unrest as well as pandemic conditions, may make it impossible to complete NRFU by September 30, 2020. This is particularly true in states with the lowest total response rates—particularly Alabama, Mississippi, Louisiana, Georgia, South Carolina, North Carolina, and Florida where weather events may make it impossible for the Census Bureau to complete NRFU operations in some Area Census Offices by September 30, 2020.
  - A shortened NRFU may increase the undercount of newborns as well as other children. Newborns, who are historically undercounted, must be enumerated through household self-response or NRFU and cannot be imputed through tax records. NRFU is an important method by which children under 5 years old are counted.
  - Groups with lower internet access, such as lower income individuals, rural residents, Native Americans, and others are at risk of being more significantly undercounted than in 2010 if NRFU is shortened. These groups are historically undercounted, but the risk of undercounting them is amplified with the pandemic, the reliance on the internet, and the shortened NRFU timeline.

• Reduced contact attempts for self-reported vacant housing units, re-interviews, and self-response quality assurances, without testing the impact on data quality, increases the risk of errors during NRFU.

# **CENSUS BUREAU RESPONSE:**

We understand the Committee's concerns about accuracy of the 2020 Census. The Nonresponse Follow-up operation concluded on October 15, 2020. All states, the District of Columbia, and the Commonwealth of Puerto Rico, achieved total completion rates over 99.0%; and all but one state was at or above 99.9%. While other quality measures for the 2020 Census will be produced over time, it helps to look at selfresponse rates as an initial metric. Our final self-response rate for the 2020 Census was 67.0%, compared with the final self-response rate for 2010 of 66.5%. Generally, better data comes from self-response, and after a decade of global decline in census and survey participation along with the challenges presented by COVID-19, we were able to beat the 2010 self-response rate.

<u>https://www.census.gov/content/dam/Census/newsroom/press-kits/2020/nrfu-</u> <u>deadline-completion-rates-faq.pdf</u>

- 3) CSAC remains concerned about the accuracy of the final 2020 Census data based on the shortened time frame for the post data collection processing operation. While previous censuses have required 5 months of post data collection processes, the 2020 Census compressed timeline will only give the Bureau 3 months to complete these tasks, with several data checking processes eliminated. CSAC advises the Census Bureau that the following issues may compromise the accuracy of the 2020 Census under the "Replan" compressed timeline for Post Data Collection Processing:
  - Given that large numbers of people are changing their normal residential patterns due to pandemic conditions (e.g., college students, snowbirds), adequate deduplication procedures for college students, retirees, and others require additional time.

# **CENSUS BUREAU RESPONSE:**

It is possible for college students to be counted by the student housing administrator during the group quarters data collection operation, as well as by administrative records if they lived in off campus housing but went home during the outbreak. Both situations could also result in parents including the student in their housing unit response to the 2020 Census. The Coverage Improvement operation will unduplicate these students if a good address is provided for where they lived while attending school. Regarding the de-duplication procedures in our post-processing systems, they are fully tested and prepared to execute every step of the process. None of the changes in schedule or duration caused by this year's challenges has led to deleting or altering any processing steps. Our software will run as intended, and we will address all data anomalies and errors, including duplicate responses.

• Elimination of expert review of group quarters by local state demographers through Count Review Event 2 increases the risk that the Census Bureau will publish data with errors in the group quarters population.

#### **CENSUS BUREAU RESPONSE:**

FSCPE members were able to provide additional information for group quarter facilities that were not enumerated during group quarters enumeration. These group quarters were contacted again during a modified late group quarters to attempt to collect demographic data for residents or a minimum population count as a last resort.

• Untested post data collection processing systems may fail in ways that the Census Bureau cannot foresee today.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau has thoroughly tested post data collection processing systems, and as stated above, we will address all data anomalies and errors. Additionally, we have been able in this Census to execute pre-processing data checks of states that were approaching completion during data collection. This has given us the ability to discover and address issues earlier. We anticipate that this work will also enable faster and more accurate processing.

In his September 11th statement, the Associate Director for Decennial Census Programs acknowledged that changes to post data collection processing procedures increase the risk of unidentified errors in the collected data.<sup>1</sup> In this CSAC meeting, many of these processes were characterized as redundant. However, redundancy in data checks is necessary to ensure the accuracy of the Census results, just as redundancy in data collection systems was essential to gather the best quality data. Some post data collection processes may seem redundant before executing them, but previous census experience has consistently shown that post data collection activities are an essential tool for ensuring the quality of Census results.

#### **CENSUS BUREAU RESPONSE:**

The Census Bureau agrees with this comment. The only redundant process we have eliminated is a review by DSSD of the Master Address File (MAF) 2020 extract We made this decision because there have been decades of work and numerous processes throughout 2020 Census preparations and operations to ensure the MAF is complete and accurate. No other redundant processing steps were eliminated. Any other changes to backend processing only included

<sup>&</sup>lt;sup>1</sup> https://assets.documentcloud.org/documents/7207428/LUPE-Sept-11-2020-Declaration-of-Albert-Fontenot.pdf

reduced calendar time. We mitigated these reductions by working around the clock and adding weekend and holiday production days to the schedule. By closely managing both the process and the risks, our intention continues to be to deliver an accurate, defensible 2020 Census count as close to the statutory deadlines as possible.

4) As a result, the CSAC recommends that the Census Bureau have the time it requested in April 2020 to execute its full battery of data checks to reduce the risk of failing to identify key errors and generate final 2020 Census products that are of comparable quality to previous decennial censuses. Specifically, CSAC believes that the Census Bureau needs the full six months it requested in April 2020 for post data collection processes.

#### **CENSUS BUREAU RESPONSE:**

Though the Census Bureau agrees that additional time can further reduce the risk of missing processing errors, we unfortunately have no control over how the U.S. Congress will proceed with legislation to increase the time available for delivery of the 2020 Census results. However, as stated above, by closely managing both the process and the risks, our intention continues to be to deliver an accurate, defensible 2020 Census count as close to the statutory deadlines as possible.

Moreover, the pandemic, natural disasters, late changes to processes, and accelerated timetables are also impacting one of the key tools for measuring decennial Census quality, the Post-Enumeration Survey (PES). Given the known challenges with the planned Post-Enumeration Survey including difficulty of recall and non-response bias, the coverage error in the 2020 Census may not be well measured by the PES. Therefore, it is even more important that the Census execute all originally planned post data collection processes as well as any additional processes envisioned in April 2020 to ensure the 2020 Census data are as accurate as possible.

#### **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts the recommendation to conduct a high-quality Post-Enumeration Survey and acknowledges that there are numerous challenges to the planned Post-Enumeration Survey. The Post-Enumeration Survey (PES) is underway. The first two field operations have concluded (Independent Listing and the Initial Housing Unit Follow up) and the third field operation, Person Interviewing, began on September 23.]

In response to these challenges, the Census Bureau formed a PES Quality Team to document challenges and impacts resulting from COVID-19 as well as to identify changes to mitigate and measure errors potential errors. Although COVID-19 has resulted in additional challenges for the PES, such as the potential for more recall error and nonresponse error, we are taking measures to mitigate and correct for many of these potential errors. In general, durations for PES operations are either unchanged or lengthened as a result of the schedule delays.

5) Lastly, to increase quality, CSAC recommends that the Bureau publish daily response rates that include self-response and NRFU completions, at the census tract level. This will support the work of partner organizations in targeting their final outreach efforts to the specific neighborhoods where response rates fall shortest of the 99 percent goal.

### **CENSUS BUREAU RESPONSE:**

The Census Bureau does not accept this recommendation for the 2020 Census. The data used to post our NRFU completion rates are not readily available during production at the census tract level. The information we have been providing at the state and ACO levels provides more transparency into the progress of our NRFU operation than in any previous census. We will carry this recommendation forward to our planning for the 2030 Census.

#### Administrative Records Use

CSAC commends the Bureau for bringing years of developmental work to fruition in the implementation of administrative records (AR) use to reduce the NRFU workload and enumerate a portion of the nonresponding households.

- 1) CSAC recommends a thorough assessment of this novel application to be presented in a public report. This assessment should include:
  - Estimated cost savings (potential visits saved, and the dollars associated with these visits)
  - Estimated accuracy
  - Model stability over 10 years
  - Ways to improve both the roster building and predictive modeling

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation. The administrative record usage will be documented in the Nonresponse Follow-up operational assessment. The Census Bureau thanks the committee for these suggestions and will attempt to address them in public reports as much as possible.

- 2) With respect to assessing accuracy, CSAC recommends consideration of the following approaches among others that the Census Bureau may propose:
  - For the set of addresses where prospective AR households were replaced by late self-responses, compare the counts and composition between the two.

• For self-responding addresses that would have qualified as AR Occupied if they had been part of the NRFU workload, compare the counts and composition between the self-responses and the AR Occupied.

# CENSUS BUREAU RESPONSE:

# The Census Bureau accepts this recommendation and will address these questions in the Nonresponse Follow-up operational assessment.

3) The accurate enumeration of college students both on-campus and off-campus was especially challenging in 2020. Lessons learned from these efforts may prove valuable in the future. CSAC recommends that the Census Bureau conduct a thorough analysis of the accuracy of enumeration of these populations, using whatever methods are available. These should include both the PES and demographic analysis.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation. This will include attempts to use the PES to evaluate non-group-quarters coverage only as one limitation will be that group quarters, such as college dorms, were out of scope for that survey.

4) Since young adults have higher mobility rates even without a pandemic, CSAC recommends that the Census Bureau explores (after 2020 Census operations) discussions with the U.S. Department of Education to include an exemption to FERPA<sup>2</sup> to allow colleges and universities to share student information for purposes of the decennial Census enumeration and/or post data collection processing (protected by Title 13 and not usable for other purposes).

#### **CENSUS BUREAU RESPONSE:**

#### The Census Bureau accepts this recommendation.

 CSAC recommends that the Census Bureau explore a partnership with universities that would facilitate data sharing to improve potential enumeration of students in university locales.
 CENSUS BUREAU RESPONSE:

#### The Census Bureau accepts this recommendation.

6) CSAC recommends that the Census Bureau explore obtaining birth records for the first three months of the year to facilitate roster building.

### **CENSUS BUREAU RESPONSE:**

<sup>&</sup>lt;sup>2</sup> Family Educational Rights and Privacy Act

The Census Bureau accepts the recommendation. The Census Bureau has already obtained the updated birth applications for the first three months of the year from the Social Security Administration. The Census Bureau will evaluate these records to see if there are ways for the 2030 Census that these records could be included.

7) CSAC recommends exploring whether the savings from AR use in NRFU can be applied to improve on-the-ground enumeration.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation. The plan to use administrative records where they were of high quality for the 2020 Census was part of the plan to improve on-the-ground enumeration. The plan that was researched and tested throughout the decade was that non-responding units with no or lower-quality administrative records could be targeted for relatively more visits as compared to the addresses with higher quality administrative records.

8) CSAC recommends exploring whether data on cell phones, given their ubiquity, can be used to improve counting of the hard-to-count populations.

# **CENSUS BUREAU RESPONSE:**

*The Census Bureau accepts this recommendation as a possibility to be researched and tested for the 2030 Census.* 

# **Differential Privacy**

CSAC commends the Bureau for recognizing and demonstrating the vulnerability of classic Disclosure Avoidance techniques. Reconstruction and re-identification risks are serious and are growing with the increase in computational power and availability of auxiliary data sets.

Census data require protection, and CSAC commends the Bureau for its serious commitment to modern and future-proof privacy protection and its development of differential privacy protocols. Further, CSAC notes that the Bureau's implementation of differential privacy at the scale of the 2020 Census via its TopDown Algorithm (TDA) is an exceptional technical achievement. In the course of developing its differential privacy algorithms and code, the Bureau used an exemplary development process, following current best practices and making new contributions to the field.

In addition, the Bureau has made extensive efforts to seek input on use cases from multiple sources, and the compilation of these use cases (especially the collection of Federal Register use cases) is an excellent resource for studying the effects of differential privacy.

CSAC recognizes that the Bureau has made improvements to transparency in its development of differential privacy protections, including its maintenance of a centralized location for updates: <u>2020 Disclosure Avoidance System Updates</u>. Some aspects of the Bureau's differential privacy efforts are less transparent.

1) CSAC recommends that the Bureau make further efforts to communicate any updates on the decision-making process for the privacy-loss budget and its allocation, and any updates on the timeline for implementation of differential privacy.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation. We recognize the importance of clear communication of the policy decisions relating to the Disclosure Avoidance System. We will communicate these decisions through our 2020 Census Data Products newsletter, on our Disclosure Avoidance System Updates website, and through the 2020 Census Memorandum Series.

2) While the Bureau has collected many important use cases, CSAC recommends that the Bureau should take substantially more time to catalog methodically the use cases of census data, including funding allocations, legal mandates, and regulatory practices, across all agencies of the federal government as well as at state and local levels.

This catalog should be publicly available and will help in selecting priority use cases for analysis (see below) and in determining the overall privacy-loss budget and its allocation for the 2020 census. This catalog should be periodically updated going forward to inform decisions about how differential privacy is applied to the American Community Survey, 2030 census, or other census-derived data. Federal-State Cooperative for Population Estimates (FSCPE) members have already begun cataloging state use cases and could be partners in this work.

In the meantime, CSAC encourages the Bureau to publish the Excel workbook summarizing the use cases collected from the Federal Register on the Census Data Products website.

#### **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation, to the extent that it is feasible within the constraints of the 2020 Census operational schedule. Recognizing the innumerable and diverse downstream uses of census data at all levels of government, in academia, business, and elsewhere, a comprehensive catalog of these use cases would be nearly impossible. Consequently, the Census Bureau has focused on summarizing and cataloging the types of use cases, with a focus on assessing how to measure data accuracy and fitness for use in these diverse contexts. These use cases are cataloged in the 2010 Demonstration Detailed Summary Metrics. As new categories of use cases are identified, they will be added to this workbook. We will also publish our summary of the use cases submitted via the Federal Register notice on our Disclosure Avoidance System Updates website, and via Regulations.gov. Additionally, the Census Bureau has published a technical assessment of data fitness-for-use for redistricting and Voting Rights Act related uses, which is currently being updated to reflect the Committee's comments. We are working with external experts to identify and catalog additional use cases for the decennial census and other censuses and surveys; we will share this information with the Committee once we are able. The Census Bureau will also engage with the Federal-State Cooperative on Population Estimates to expand this catalog, especially for uses of the Demographic and Housing Characteristics files and subsequent data products.

Additional rigorous analysis is needed for different use cases, particularly analyses of impacts on funding formulas for federal agencies and Congressional staffers, and analyses of impacts on legal mandates and regulatory practices, including protections for civil rights.

 CSAC therefore recommends that the Bureau conduct analyses of the impact of differential privacy for priority use cases (funding, legal, and regulatory at all levels of government). An example of such analysis (for redistricting) is the paper "Variability Assessment of Data Treated by the TopDown Algorithm for Redistricting" (Wright and Irimata 2020).

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation. However, given the diversity and extent of these uses, a comprehensive analysis would not be possible before key decisions on the privacy-loss budget for the 2020 Census would need to be made. Consequently, the use cases represented in the 2010 Demonstration Detailed Summary Metrics are intended to be representative of the broader array of uses of census data. The Census Bureau also believes that these representative use cases, which include a substantial focus on fitness-for-use of the data for the Census Bureau's population estimates program, are highly correlated with key external demographic use cases. The Census Bureau has invested substantial staff resources in performing these analyses but must also rely on our data users to supplement these efforts. Of particular note, we have partnered with a group of experts identified by the Committee on National Statistics to evaluate and improve the ways we are evaluating fitness-for-use of these data. Our public release of the Detailed Summary Metrics and Privacy-Protected Microdata Files are also intended to allow external data users to perform their own analyses to supplement our own internal evaluations. Feedback from these analyses can be submitted to 2020DAS@census.gov. Additionally, the Census Bureau has published a technical assessment of data fitness-for-use for redistricting and Voting Rights Act related uses, which is currently being updated to reflect the Committee's comments.

2) For example, CSAC recommends a careful study of the impact of Differential Privacy (DP) on the Population Estimates program data, which are used for planning purposes and as an input for other data like the American Community Survey. Using the Fall 2019 demonstration data, the differences between DP version and SF1 version of these base data are large.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation and acknowledges that evaluations of the data's fitness-for-use for the population estimates program and other downstream internal uses are ongoing.

CSAC appreciates the Bureau's efforts in creating the 2010 Demonstration Products, the Sprint II Detailed Summary Metrics and other updates, and the privacy-protected microdata for evaluation by the community of users. Metrics are essential for users to judge the quality and fitness for use of Census data products. The Bureau has developed, computed, and released a set of useful metrics based on the privacy-protected 2010 Census data. These published metrics were instrumental in helping the community of users to recognize problems with the October 2019 release of the 2010 demonstration products. CSAC applauds the Bureau for adapting its algorithms in response to feedback from that community.

3) While the set of published metrics is very useful, CSAC recommends that the Bureau publish further details on some variables (e.g., housing vacancy status - seasonal homes) and that some geographies should be included/better represented (e.g., zip codes, county subdivisions/minor civil divisions). As another example, the Bureau should analyze how aggregating data from small geographic units affects accuracy.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation. Many of these new disaggregations and new geographic levels are already included in the revised version of the Detailed Summary Metrics (e.g., school districts, MCDs). These additions and revisions are based on feedback we have received from our data users. The Census Bureau welcomes the Committee's suggestions for additional revisions and will conduct an assessment of including them in future versions of the metrics on an ongoing basis.

4) The recommended use case catalog development and rigorous analysis for priority use cases may suggest the need for new metrics, in addition to those metrics that have already been developed. CSAC recommends that the Bureau revisit the list of metrics periodically as the use case catalog and analyses evolve, to see if additional kinds of metrics would be useful.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation. A variety of new use cases, supplied by our data users, are already included in the revised version of the Detailed Summary Metrics. We will continue to expand and update these metrics as new representative use cases are identified.

5) CSAC appreciates the value of the privacy-protected microdata for evaluating data quality but use of these data is challenging even for sophisticated users. To aid further in the assessment of the quality of the privacy-protected data, CSAC recommends that the Bureau release additional versions of the Detailed Summary Metrics, including quality metrics at a finer scale than the current overall means, by releasing means within bins. For example, the current MALPE (Mean Algebraic Percent Error) statistic could be split into the average negative relative error and the average positive relative error, rather than combining the two. Other statistics might be split at scientifically meaningful thresholds or at variable-specific cut points, like the quintiles of the distribution.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau thanks the committee for this recommendation. There is no longer sufficient staff time to incorporate general recommendations into the Detailed Summary Metrics. However, we would consider adding a metric if the committee identified a specific use case not covered by the current metrics. The Census Bureau will continue releasing periodic Privacy-Protected Microdata Files (PPMF) to support external users who wish to expand the metrics.

6) The post-processing within the TopDown Algorithm (TDA) can create positive biases, particularly in small domains where rounding up occurs to avoid negative values. A concern is that these small positive biases can accumulate as small domains are combined to create custom geographies. To facilitate assessment of bias properties for the privacy-protected data, CSAC recommends that the Bureau should release the non-post-processed data used in TDA, which are unbiased estimates with known error distributions.

# **CENSUS BUREAU RESPONSE:**

The Census Bureau accepts this recommendation and is currently investigating how to provide this guidance. In order to be a properly supported 2020 Census data product, the Census Bureau will need to develop a set of tools to inform our data users on how to use and interpret the data. We will supply this guidance, though it may not be released with the publication of the official 2020 Census data products and will not be available for the 2010 Demonstration Privacy-Protected Microdata Files. As final decisions are made, we will communicate to the public.

7) The Bureau should make clear what, if any, metrics for 2020 will be computed from 2020 data. The Bureau should make readily available tools for extrapolating from 2010 demonstration metrics to 2020 use cases. A specific suggestion for such a tool is for the Bureau to develop "Generalized Metrics Functions (GMFs)" by analogy to Generalized Variance Functions. A GMF would be obtained by regression of 2010 metrics on 2010 privacy-protected tabular estimates and cell sizes. The fitted regression model could then be used to estimate 2020 metrics, by plugging in 2020 privacy-protected tabular estimates and cell sizes.

# **CENSUS BUREAU RESPONSE:**

*The Census Bureau accepts this recommendation. We are currently researching how to produce and provide these tools.* 

CSAC has been asked to advise on prioritization of use cases in the allocation of the privacy-loss budget (PLB) across data products. Due to the complexities of the disclosure-avoidance system, the implications of the PLB allocation for privacy, for accuracy, and for the privacyaccuracy trade-off are unclear. CSAC is not aware of either theory or empirical data sets that would offer guidance in addressing these questions. The Bureau may be required to produce, from administrative records, estimates of undocumented individuals counted in the 2020 Census, for December release with the state apportionment counts, and the Bureau is developing estimates of the number of citizens in each block based on administrative records for CVAP for release in 2021.

- 8) Given the use cases that CSAC has considered, and the committee's assessment of potentially missing use cases, CSAC recommends that the privacy-loss budget should be prioritized toward the most important use cases in this order:
  - Government funding (federal, state, local)
  - Legal mandates and regulations
  - Community planning (children's & elder services, infrastructure)

# **CENSUS BUREAU RESPONSE:**

The Census Bureau thanks the Committee for this recommendation and will take this under advisement. Our primary focus, to date, has been improving accuracy in population totals for political entities, which largely supports the government funding priority that the Committee has identified. Our secondary priority on improving accuracy for the Census Bureau's population estimates program also supports the Committee's proposed priorities. That said, prioritizing accuracy for population and demographic characteristics for arbitrary geographic entities (e.g., the redistricting use case) is challenging because these geographic areas cannot be defined in advance. This is why the Census Bureau has used the principle that the accuracy of any statistic produced by the DAS must increase as the population in the category or geography increases.

The Bureau may be required to produce, from administrative records, estimates of undocumented individuals counted in the 2020 Census, for December release with the state apportionment counts, and the Bureau is developing estimates of the number of citizens in each block based on administrative records for CVAP for release in 2021. In either case, citizenship status would receive a share of the privacy-loss budget and would reduce the accuracy and usability of other variables.

9) CSAC recommends that if any citizenship variables are part of the December release or CVAP release, the Bureau should assign to these variables a very small part of the privacy-loss budget, such that these data will be more protected. The citizenship data are more sensitive than many other attributes. This attribute is much more correlated within geographic locations, making reconstruction attacks on the data that account for such correlations much more effective in recovering this attribute. The imputations from the Census Unedited File (CUF) to the Census Edited File (CEF) increase the impact of any one person's data on the output, and thus increase the privacy leakage through this attribute. Further, given the inherently large uncertainties in the imputed citizenship attribute, it may be more beneficial to place more of the privacy loss budget on the more accurate tabulations.

# CENSUS BUREAU RESPONSE:

The Census Bureau thanks the Committee for this recommendation and will take it under consideration. The Census Bureau's Data Stewardship Executive Policy Committee (DSEP) will make final determinations on the level of privacy protection for these data.

The Bureau's implementation of differential privacy has followed an ambitious timeline under any circumstances, even in the absence of a global pandemic or other challenges. The Bureau is operating under enormous time pressure to make the incredibly consequential and irreversible decision on the privacy-loss budget and its allocation. But many implications of this decision for privacy, accuracy, and fitness-for-use are currently unknown. The process by which the Bureau will determine the privacy-loss budget allocation is unclear. Whatever the choice of privacy-loss budget allocation, the Bureau will need to estimate the re-identification risk to ensure sufficient privacy, will need to give users methods for assessing fitness-for-use, and will need to have a backup plan (e.g., allocate some privacy budget) for the future, in case differentially-private data are not fit for some important use cases. The recommended use case catalog development and rigorous analysis for priority use cases are important for informing how to allocate the privacy-loss budget across uses. 10) CSAC recommends that the Bureau should delay additional releases after the December apportionment release to allow time for these recommended analyses.

#### **CENSUS BUREAU RESPONSE:**

The Census Bureau thanks the Committee for this recommendation and will take it under advisement. Public Law 94-171 requires the Census Bureau to produce and deliver the redistricting data to the states, the District of Columbia, and Puerto Rico within one year following Census Day. Therefore, we are focusing our attention to identifying the activities and schedules needed to attempt to meet this requirement.

Planning for the production and release of the remaining 2020 Census data products will restart immediately following completion of the redistricting data planning activities.

#### **Public Comments**

CSAC appreciates the Census Bureau enabling public engagement and recommends that the Census Bureau respond in writing to the four written public comments.

#### **CENSUS BUREAU RESPONSE:**

Thank you for your recommendations.

Case 3:21-cv-00211-RAH-ECM-KCN Document 94-14 Filed 04/20/21 Page 1 of 3

# Exhibit 14

 To:
 Danah\_Michele\_Boyd (CENSUS/CED\_CTR)[danah.m.boyd@census.gov]
 Filed\_04/20/21
 Page 2 of 3

 From:
 John Maron Abowd (CENSUS/ADRM FED)[/0=ExchangeLabs/ou=Exchange Administrative Group
 Filed\_04/20/21
 Page 2 of 3

 (FYDIBOHF23SPDLT)/cn=Recipients/cn=cb0eee1cc6ca45cc948c0077899626c2-Abowd, John]
 Sent:
 Wed 10/7/2020 2:44:44 PM (UTC)

 Subject:
 Re: Letter to DSEP committee
 Sent:
 Ne: Letter to DSEP committee

OK. Thanks and sorry for the delayed response. This why the Data Quality EGG wants a CNSTAT consensus panel with unfettered access to the raw data. And why I hired Paul Biemer to do a total uncertainty analysis of the 2020 Census. But I don't know how to speed up the work. I agree with Tori's assessment that we should hold everything after the redistricting data until there is consensus that the underlying data are fit-for-use. But there's no way to slow the redistricting long enough to get that. Let's talk. But not today.

Thanks,

John M. Abowd, PhD, Associate Director and Chief Scientist Research and Methodology U.S. Census Bureau O: <u>301-763-5880</u> M: simulring on cell <u>census.gov</u> | <u>@uscensusbureau</u> Shape your future. START HERE > 2020census.gov

From: Danah Michele Boyd (CENSUS/CED CTR) <danah.m.boyd@census.gov> Sent: Monday, October 5, 2020 4:47 PM To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov> Subject: Re: Letter to DSEP committee

The way that I read this letter echoes what I've been hearing from other stakeholders. Q1, which is rooting all other questions right now, is: Will decennial data (as in the CEF) be fit-for-use? The Census Bureau is likely to say yes, but will data users have confidence in that vocalization? Data users want to be able to verify that assertion. In the last, they used the published data to get a sense for whether there were problems because they felt as though, even with swapping or other disclosure avoidance components, they could detect fundamental errors in field operations through the published data.

While concerns about underlying data are not new, they are unquestionably higher because of COVID. So data users want to understand if they'll even know where the problems might be.

This then factors into the DAS. Q2: Will identified problems be rooted in the DAS or the CEF (aka decennial and field)? No one has the foggiest clue how to disentangle those two components on the outside.

Given the high level of uncertainty about the factors that are influencing the data in the CEF and the expectation that there will be challenges to the CEF, more folks are asking for the DAS to take a backseat. This poses a fundamental question for you: How can the CEF data be meaningfully audited/verified such that external stakeholders have confidence in the CEF (or understand its limitations) even if they cannot directly access/analyze it? As long as these two facets are entangled, the DAS will take the blame for any problems in the 2020 census data because it's what the public has access to. (I am dreading public comments like "the best census ever!" from Bureau officials because these will be the hook.)

In terms of how to respond, I would probably break down the puzzle being articulated so that he knows you're hearing him. He's looking for the data to be validate-able, for it to not cause XYZ ripple effects, for errors in the DAS to be discernable. And then talk through approaches for achieving those goals AND ensuring T13 compliance. "You have to trust us" ain't gonna work.

Happy to talk through this cuz I think this is deeper than the DAS. The DAS is the tip of the data quality iceberg. This census is triggering deep epistemological questions: how do we know what we know? And that's the funny thing about methodological transparency - most folks don't feel reassured by understanding how knowledge is produced; only scientists do. This is why there's a long history of data being \*performed\*. By trying to explain, you built a lightning rod.

But the DAS isn't the metal at the basen it's wet the rod. And there's not hat's graved in the tod sight now. The big question is whether or not that rod can be grounded.

Let me know if you want to strategize.

**danah boyd**, Research Contractor Research and Methodology U.S. Census Bureau M: <u>202.860.4551</u>

census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov> Sent: Monday, October 5, 2020 2:30 PM To: Danah Michele Boyd (CENSUS/CED CTR) <danah.m.boyd@census.gov> Subject: Fw: Letter to DSEP committee

Suggestions welcome. He is typical of FSCPE. Thanks,

John M. Abowd, PhD, Associate Director and Chief Scientist Research and Methodology U.S. Census Bureau O: <u>301-763-5880</u> M: simulring on cell <u>census.gov</u> | @uscensusbureau Shape your future. START HERE > <u>2020census.gov</u>

From: Jan Kasper Vink <jkv3@cornell.edu>
Sent: Monday, October 5, 2020 2:23 PM
To: Census 2020 DAS <2020das@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED)
<cynthia.davis.hollingsworth@census.gov>; John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Subject: Letter to DSEP committee

It is my understanding that later this week the DSEP committee will discuss development and invariants of the Disclosure Avoidance System.

I have deep concerns about the TDA system that is currently in development and hope that attached document with these concerns finds its way to the members of the DSEP committee for consideration.

With best regards,

Jan Vink Extension Associate Cornell Program on Applied Demographics Web: <u>https://pad.human.cornell.edu</u>, Twitter: @PADCornell Tel: (607) 255 8399