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

THE STATE OF ALABAMA; *et al*., Plaintiffs, v. UNITED STATES DEPARTMENT OF COMMERCE; *et al.*, Defendants.	CIVIL ACTION NO. 3:21-CV-211-RAH

PLAINTIFFS' NOTICE OF EVIDENTIARY SUBMISSION

Plaintiffs hereby submit the attached supplemental exhibits pursuant to this Court's Order of April 19, 2021. *See* (ECF No. 93). The attached supplemental exhibits are in addition to the evidence Plaintiffs have already submitted with their prior pleadings. All exhibits, including but not limited to those previously filed and attached to Plaintiffs' Motion for Preliminary Injunction, Plaintiffs' Reply in Support of their Request for Three-Judge Court, and Plaintiffs' Reply in Support of their Preliminary Injunction, are hereby included herein as exhibits by reference. *See* (ECF Nos. 3, 25, & 94).

Plaintiffs submit the following list of exhibits for the Court's consideration at the hearing on the preliminary injunction set for May 3, 2021, reserving the right to supplement, modify, or amend this list to include material which may be needed to rebut evidence presented by the Defendants. Plaintiffs further reserve the right to modify this list with leave of Court.

Plaintiffs' Exhibit List

Exhibit	Exhibit Name
PL EX 1	Simson L. Garfinkel, John M. Abowd, & Sarah Powazek, Issues Encountered Deploying Differential Privacy 3 (Sept. 6, 2018), https://arxiv.org/pdf/1809.02201.pdf

PL EX 2	U.S. Census Bureau, 2020 Disclosure Avoidance System Updates (Feb. 23, 2021), https://www.census.gov/newsroom/press-releases/2020/statement-covid-19-2020.html
PL EX 3	U.S. Department of Commerce Secretary Wilbur Ross and U.S. Census Bureau Director Steven Dillingham Statement on 2020 Census Operational Adjustments Due to COVID-19 U.S. Census Bureau (Apr. 13, 2020), https://www.census.gov/newsroom/press-releases/2020/statement-covid-19-2020.html
PL EX 4	2020 Census Response Rate Update: 99.98% Complete Nationwide, U.S. CENSUS BUREAU (Oct. 19, 2020), https://www.census.gov/newsroom/press-releases/2020/2020-census-all-states-top-99-percent.html#:~:text=OCT.,15%2C%202020.
PL EX 5	Albert E. Fontenot, 2020 Census Update, Presentation to the Census Scientific Advisory Committee March 18, 2021 at 2, https://www2.census.gov/about/partners/cac/sac/meetings/2021-03/presentation-2020-census-operational-review.pdf
PL EX 6	Michael B. Hawes, U.S. Census Bureau, Implementing Differential Privacy: Seven Lessons From the 2020 United States Census, Harvard Data Science Review (Apr. 30, 2020) https://hdsr.mitpress.mit.edu/pub/dgg03vo6/release/2
PL EX 7	Michael Hawes, U.S. Census Bureau, Differential Privacy and the 2020 Decennial Census at 24 (Mar. 5, 2020), https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf
PL EX 8	John M. Abowd, Modernizing Disclosure Avoidance: A Multipass Solution to Post-processing. Error, The Census Bureau, (June 18, 2020), https://www.census.gov/newsroom/blogs/research-matters/2020/06/modernizing_disclosu.html
PL EX 9	Andy Beveridge, Sixteen States Sue to Block Census Bureau Data Privacy Method (Apr. 19, 2021), https://www.socialexplorer.com/blog/post/sixteen-states-sue-to-block-census-bureau-dataprivacy-method-11411
PL EX 10	Aref N. Dajani et al., Presentation to Census Scientific Advisory Committee, The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau (Sept. 2017) https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf
PL EX 11	John M. Abowd et al., The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau (July 2020) https://www2.census.gov/adrm/CED/Papers/CY20/2020-08-AbowdBenedettoGarfinkelDahletal-The%20modernization%20of.pdf
PL EX 12	U.S. Census Bureau, Decennial Census P.L. 94-171 Redistricting Data (Mar. 15, 2021) https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html

PL EX 13	U.S. Census Bureau, Meeting Redistricting Data Requirements: Accuracy Targets (Apr. 7, 2021), https://content.govdelivery.com/accounts/USCENSUS/bulletins/2cb745b
PL EX 14	Amy Lauger et al., U.S. Census Bureau, Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research 2 (Sept. 26, 2014), https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/cdar2014-02-discl-avoid-techniques.pdf
PL EX 15	JASON, Formal Privacy Methods for the 2020 Census (Apr. 2020), https://www.census.gov/programs-surveys/decennial-census/2020-census/planningmanagement/planning-docs/privacy-methods-2020-census.html
PL EX 16	Laura McKenna, U.S. Census Bureau, Research & Methodology Directorate: Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Census of Population and Housing Public Use Microdata Samples (Apr. 2019) https://www2.census.gov/adrm/CED/Papers/FY20/2019-04-McKennaSix%20Decennial%20Censuses.pdf
PL EX 17	10.22.2020 Cogley to Mayfield For Deputy Directors Approval
PL EX 18	10.7.2020 Disclosure Avoidance System Power Point
PL EX 19	11.14.2019 Abowd Power Point Census Bureau Good Data Steward
PL EX 20	2.26.20 Letter from J Abowd to Steering Committee
PL EX 21	4.10.20 Concerns about Disclosure Avoidance Program Letter from C Benson to S. Dillingham
PL EX 22	5.22.2020 Reducing the Magnitude of Upward Bias
PL EX 23	6.26.2020 Thieme to Adams File Integrity Checks
PL EX 24	7.13.2020 Abowd to Hill Question on DAS
PL EX 25	7.21.2020 Abowd to Velkoff Large Epsilon Additional Runs for Review
PL EX 26	7.21.2020 Hollingsworth to Velkoff Large Epsilon Additional Runs for Review
PL EX 27	7.24.2020 Hawes to Abowd Fact Checking Request
PL EX 28	7.27 to 7.31.2020 Annual Conference Power Point
PL EX 29	7.9.2020 Ingold to Whitehome Silver Lining Decades Ago
PL EX 30	8.19.2020 Philip LeClerc The Challenge of Invariants and the Microdata Requirement
PL EX 31	9.18.2020 Allison Plyer to Dillingham Recommendations and Comments to the Census Bureau
PL EX 32	9.18.2020 Plyer to Dillingham Recommendations and Comments to the Census Bureau
PL EX 33	Draft of Memo about Concerns with Intentionally Distorting the Population Tabulations
PL EX 34	FY 2021 DOC Senate QFRs Document
PL EX 35	6.24.2020 Abowd Letter to Steering Committee
PL EX 36	Expert Report of Michael Barber in Reply to Amicus Brief of Data Privacy Experts
PL Demonstrative EX 1	Census and Redistricting Overview Presentation

Dated: April 26, 2021

Respectfully submitted,

STEVE MARSHALL
Attorney General of Alabama

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

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

James W. Davis (ASB-4063-I58J)
Winfield J. Sinclair (ASB-1750-S81W)
Brenton M. Smith (ASB-1656-X27G)
Assistant Attorneys General

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

Counsel for the State of Alabama

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

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

*admitted *pro hac vice*

Counsel for Plaintiffs

CERTIFICATE OF SERVICE

I hereby certify that on April 26, 2021, I served a copy of the foregoing by CM/ECF to all counsel of record.

/s/Jason Torchinsky
Jason Torchinsky
VA Bar No. 47481
15405 John Marshall Hwy
Haymarket, VA 20169
P: (540) 341-8808
F: (540) 341-8809
E: JTorchinsky@hvjt.law
Counsel for Plaintiffs

EXHIBIT 1

Simson L. Garfinkel, John M. Abowd, & Sarah Powazek, Issues Encountered Deploying Differential Privacy 3 (Sept. 6, 2018),

https://arxiv.org/pdf/1809.02201.pdf

Issues Encountered Deploying Differential Privacy

Simson L. Garfinkel
US Census Bureau
Suitland, MD
simson.l.garfinkel@census.gov

John M. Abowd
US Census Bureau
Suitland, MD
john.maron.abowd@census.gov

Sarah Powazek
MIT
Cambridge, MA
powazek@mit.edu

ABSTRACT

When differential privacy was created more than a decade ago, the motivating example was statistics published by an official statistics agency. In attempting to transition differential privacy from the academy to practice, the U.S. Census Bureau has encountered many challenges unanticipated by differential privacy's creators. These challenges include obtaining qualified personnel and a suitable computing environment, the difficulty accounting for all uses of the confidential data, the lack of release mechanisms that align with the needs of data users, the expectation on the part of data users that they will have access to micro-data, and the difficulty in setting the value of the privacy-loss parameter, ϵ (epsilon), and the lack of tools and trained individuals to verify the correctness of differential privacy implementations.

CCS CONCEPTS

• **Security and privacy → Privacy protections**; • **Theory of computation → Theory of database privacy and security**; • **Software and its engineering → Software verification**;

KEYWORDS

Differential privacy, US Census Bureau

ACM Reference Format:
Simson L. Garfinkel, John M. Abowd, and Sarah Powazek. 2018. Issues Encountered Deploying Differential Privacy. In *2018 Workshop on Privacy in the Electronic Society (WPES'18), October 15, 2018, Toronto, ON, Canada,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3267323.3268949

1 INTRODUCTION

The U.S. Census Bureau is the largest agency in the Federal Statistical System. According to the Census Bureau's mission statement, "The Census Bureau's mission is to serve as the leading source of quality data about the nation's people and economy. We honor privacy, protect confidentiality, share our expertise globally, and conduct our work openly."[22]

As the 2020 Census approaches, focus turns to the Census Bureau as it deploys differential privacy to protect privacy in the upcoming decennial census. Invented by Dwork et. al in 2006, differential privacy provides a mathematical definition for the privacy loss to individuals associated with the publishing of statistics based on their confidential data. Today the differential privacy literature provides numerous mechanisms for privacy preserving data publishing and privacy preserving data mining while limiting the resulting privacy loss to mathematically provable bounds[11].

The 2020 Census data processing system begins by attempting to collect data from all people living in the United States through a variety of means, including an online instrument, a telephone voice-response system, a form that can be mailed in, and "enumerators" who travel from house-to-house for non-response follow-up (NRFU)[21]. These confidential data will collected and processed to create the Census Unedited File (CUF), which will contain a block-by-block list of every person in the United States. These data must be completed in time to meet the statutory deadline for reapportioning the House of Representatives (December 31, 2020). Subject matter experts working with Census-developed software review the CUF and make corrections based on their expertise and other data sources. The result is the Census Edited File (CEF). The Disclosure Avoidance System (DAS), currently under development, will use a novel differential privacy mechanism to add noise to the CEF, producing the Microdata Detail File (MDF) that the Census Bureau's tabulation system will use to create the traditional data products.

In 2008, the Census Bureau deployed OnTheMap, the first production system to use differential privacy[6]. Six years later, Google deployed RAPPOR[12], the second major production system to use differential privacy, in its Chrome web browser. Today, differential privacy is also being used by Apple[7] and Microsoft[9]. Although these examples all use differential privacy to protect data supplied by individuals, they use it in different ways, for different purposes. The Census Bureau operates as a *trusted curator*, which collects sensitive data from individuals, performs statistical tabulations, and publishes them. Trusted curators use differential privacy to prevent matching between a respondent's identity, their data, and a specific data release, which is the Census Bureau's legal requirement under Section 9 of the Census Act, U.S. Code Title 13. Google, Apple and Microsoft use the *local model* of differential privacy: randomization is performed by software running on the individual's computer. These companies use differential privacy so that they cannot make reliable inferences about specific users. These companies use differential privacy to increase public acceptance of their data collection methods.

In 2017, the Census Bureau announced that it would be using differential privacy as the privacy protection mechanism for the 2020 Census of Population of Housing[14]. There is no off-the-shelf mechanism for applying differential privacy to a national census. Although in principle, the Census Bureau could apply Google's RAPPOR mechanism to the raw census returns, any resulting tabulations would contain far too much noise for any sensible value

WPES'18, October 15, 2018, Toronto, ON, Canada

ACM ISBN 978-1-4503-5989-4/18/10…$15.00
https://doi.org/10.1145/3267323.3268949

of ϵ to be of much statistical value. To use the Census Bureau's terminology, the resulting statistics would likely not meet "fitness for use" standards, which are also part of the mandate in the Census Act. The same result would ensue if the Census Bureau employed the original Laplace Mechanism[10] to protect its publication tables. An added complication of the Laplace Mechanism is that the tables would not be internally consistent, which might create concerns for data users. Instead, the Census Bureau revealed that it was developing, implementing, testing, and deploying a new differential privacy mechanism. It committed to publishing the mechanism in the peer-reviewed academic literature and making the implementation available to the public, along with suitable test data.

Surprisingly, the Census Bureau's experience with OnTheMap did not significantly prepare the organization for the difficulty of deploying differential privacy for the 2020 Census. OnTheMap was a new product that was designed to incorporate modern privacy protection. In comparison, the decennial Census of Population and Housing, first performed under the direction of Thomas Jefferson in 1790, is the oldest and most expensive statistical undertaking of the U.S. government. Transitioning existing data products to differential privacy has revealed both today's limits in the field of formal privacy, and demonstrated the difficulty of retrofitting legacy statistical products to conform with modern privacy practice.

2 PRIOR WORK

Statistical agencies of the U.S. government have traditionally used statistical disclosure limitation techniques[13] to protect confidentiality; Lauger et al. details how those techniques were applied to many data products released by the U.S. Census Bureau[15].

Abowd identifies the challenges faced by statistical agencies in reconciling their traditional disclosure limitation practices with the modern realities of database reconstruction[2, 3], which is made possible because of the large number of statistics published by official statistical agencies, the availability of large scale computational resources, and third-party data that can improve the accuracy of the reconstructed database when used in a re-identification attack.

Abowd and Schmutte proposes an approach that statistical agencies can use to set ϵ using economic theory[4].

3 SPECIFIC CHALLENGES

Here we present some of the challenges that the Census Bureau has encountered during the deployment of differential privacy. We group the challenges into those that arise from current limitations in the mathematics of differential privacy, those resulting from operational complexities within the Census Bureau, and issues faced by the agency's data users.

3.1 Scientific Issues

Differential privacy is less than 15 years old, and most existing mechanisms were created for computer science applications, not the needs of official statistical agencies.

Hierarchical Mechanisms. For the 2020 Census, the agency desired a mechanism that controlled the error as statistics were reported from smaller geographies (e.g. blocks and block groups) to larger geographies (e.g. census tracts and counties) such that the error would decrease as the population in the relevant geography increased. This required the Census Bureau to develop a set of novel hierarchical mechanisms designed to optimize the accuracy of multiple queries simultaneously.

Invariants. For the 2018 End-to-End test, the Census Bureau is reporting exact counts for some statistics (e.g. the number of people living on each block) but privatized counts for other statistics (e.g. the number of Hispanics living on each block). The agency has adopted the term *invariants* to describe statistics that are not changed by the application of differential privacy, and views them as restrictions on the universe of neighboring databases. Nevertheless, there is no well-developed theory for how differential privacy operates in the presence of such invariants. In addition, the historical reasons for having invariants may no longer be consistent with the Census Bureau's confidentiality mandate.

Stratified sampling. Between 1940 and 2000, the Census Bureau used two census forms: a *short-form* sent to the majority of households, and a *long-form* with more questions that was sent to a subset. In 2005 the Bureau replaced the long-form with the American Community Survey, a project that continuously measures the U.S. population using a stratified probability sample. Currently, there is no accepted mechanism for applying differential privacy to the results of such sampling. This has delayed the introduction of formal privacy mechanisms for the American Community Survey.

Quality Metrics. While the trade-off between between statistical accuracy and privacy loss is at the heart of differential privacy, there are many metrics for assessing the quality of a published dataset. One approach is to calculate the L_1 error between the true data (i.e. without disclosure limitation) and the privatized data. This is a coarse measure: a disclosure limited product with a high L_1 compared to the same product without disclosure limitation may still be very accurate for its intended use. Ideally, if intended uses are known, they can be incorporated into the privacy mechanism so that the usefulness is higher for the same privacy-loss budget, allowing the overall privacy-loss budget to be better deployed.

Presenting and Resolving Equity Issues. Because the Census Bureau intends to publish many tables drawn from the same confidential database and controlled by an overall privacy-loss budget, there is an opportunity to make some tables more accurate at the cost of making other tables less accurate. These can be thought of as issues of fairness between different consumers of the Census data, which can be described as an *equity issue*. In principle, these issues are no different from the decisions that statistical agencies routinely make about allocating a fixed dollar sampling budget among sub-populations in order to obtain estimates that are fit for use on those sub-populations. Differential privacy lacks a well-developed theory for measuring the relative impact of added noise on the utility of different data products, tuning equity trade-offs, and presenting the impact of such decisions.

Establishing a Value of Epsilon. Before the arrival of differential privacy at the Census Bureau, disclosure avoidance had aspects of the black arts. Knowledge of the actual disclosure avoidance techniques and parameters was restricted to a small group of specialists, and the remainder of the agency treated disclosure avoidance as a black box that input dangerous data and output clean, safe data.

The proponents of differential privacy, in contrast, have always maintained that the setting of ϵ is a policy question, not a technical one. When the Census Bureau announced that it was adopting differential privacy, it also stated that the value of ϵ would be set by policy makers, not technologists. But how should policy makers do that? Here, the literature of differential privacy is very sparse.

To date, the Census Bureau's Data Stewardship Executive Policy committee (DSEP) has set the values of ϵ for one data product. The value was set by having the practitioner prepare a set of graphs that showed the trade-off between privacy loss (ϵ) and accuracy. The group then picked a value of ϵ that allowed for sufficient accuracy, then tripled it, so that the the researchers would be able to make several additional releases with the same data set without having to return to DSEP to get additional privacy-loss budget. The value of ϵ that was given out was far higher than those envisioned by the creators of differential privacy. (In their contemporaneous writings, differential privacy's creators clearly imply that they expected values of ϵ that were "much less than one."[17]).

Mechanism Development. More efficient mechanisms and proofs with tighter bounds are needed to lower amounts of noise for the same level of privacy loss, and to make efficient use of the privacy-loss budget for iterative releases of edited and corrected statistics.

3.2 Operational Issues

Obtaining Qualified Personnel and Tools. An early problem faced by the Census Bureau was not technical, but operational: it lacked subject matter experts skilled in the theory and practice of differential privacy. In part, this is because only a smattering of universities cover the topic of differential privacy in an instructional setting, and then typically only in a single upper-level computer science course. The Census Bureau, in contrast, typically hires graduates with degrees in mathematics, statistics or economics for its "mathematical statistician" career tract. And while there is a now a textbook that covers the theory of differential privacy[11], reading a textbook does not provide the necessary expertise to develop correct differential privacy algorithms and implementations. The sparsity of expertise was noted by the Bipartisan Commission on Evidence-based Policymaking, which strongly recommended the adoption of privacy-enhancing data analysis tools while recognizing that there was a dearth of existing tools [5].

Likewise, there is a profound lack of toolkits for performing differential privacy calculations and for verifying the correctness of specific implementations. It is now 12 years since the invention of differential privacy: the situation is analogous to the state of Public Key Cryptography in 1989. This has impacted both high-profile projects such as the 2020 Census, as well as the day-to-day work involved in producing more than 100 regular data products and supporting hundreds of researchers in the Federal Statistical Research Data Centers.

Recasting high-sensitivity queries. The 2010 Census publications included statistics about individuals, statistics about households, and statistics reflecting the interaction of the two. The sensitivity of most counting queries is 1—for example, a statistic that reports the number of males and females on a block, or the number of households. Some queries that combine these kinds of statistics also have a sensitivity of 1, such as the number of households headed by a female. But some queries have a much higher sensitivity. For example, a query asking the number of children in households headed by a female has a sensitivity equal to the largest permissible household size. An added complication is that this value needs to be specified in advance, as part of the overall design of the survey, rather than derived by looking at the data, lest information about the presence of a specific large family in the survey data be revealed.

Currently, the DAS team is working with data users to redesign the publication tables, with the hope of lowering their sensitivity. For example, instead of reporting the number of children that are in a household headed by a person who is Hispanic, the Census could report the number of Hispanic children. It could also protect the original query, but at more aggregated levels of geography.

Structural Zeros. Bringing differential privacy to the 2020 Census required in-depth discussions of the difference between *structural zeros* and *sampling zeros*[8]. Structural zeros are those enforced by the Census Bureau's edit rules ("there can be no six-year old mothers with 30-year-old children"), while sampling zeros emerge from the data collection effort ("no women over 65 were found living in this facility"). Injected statistical noise can make sampling zeros positive (2 women over 65 are reported living in the facility), but cannot be allowed to undo the edit rules.

In practice, the distinction between structural and sampling zeros in an operational context is far less clear. For example, is the number of females in a male prison zero because there are none living there (a sampling zero), or because they are prohibited from living there (a structural zero)? For that matter, how should the Census determine that a facility is single-sex? Previously, whether or not a group quarters was a single-sex might have been determined by looking at the data; this is not permissible in a system that implements differential privacy.

Obtaining a Suitable Computing Environment. The algorithms being developed for the 2020 Census require significant post-processing following the application of noise. In order to characterize their behavior, Census Bureau researchers will perform many runs on the algorithms with historical data, requiring at least three orders of magnitude greater computing resources than were needed for the 2010 Census. Although the Bureau is migrating from on-premise computing to a cloud-based environment, this migration was delayed because of security concerns, resulting in substantial delays in the development of the 2020 DAS.

Accounting for All Uses of Confidential Data. A key feature of the previous disclosure avoidance mechanism was that it did not change the values of many tabulations at high levels of geography. Thus, many reports from the 2000 and 2010 Censuses could be produced using the confidential data and without applying further disclosure avoidance.

A fundamental requirement of differential privacy is that all calculations involving private data must have noise added before they

can be made public. As a result, the Census Bureau has had to identify every use of confidential data in the execution and processing of the 2020 Census. New and unanticipated requirements have emerged during the design of the system after the team thought that the design was locked down.

Lack of Final Specifications. Beyond those issues arising from the application of differential privacy, the team building the DAS has also faced by the fluid nature of the decennial census. Many of the Census Bureau data products have been traditionally developed near the end of the decade in consultation with the data users. This collaborative process helps ensure the utility of the census data, but it is at odds with the design and development of a differential privacy system, which requires that all computations be known in advance, or that some amount of privacy-loss budget be reserved for future use.

3.3 Issues Faced by Data Users

Access to Micro-data. Many Census Bureau data users are accustomed to using micro-data, like those originally released for the 1960 Census, that are either raw or that have undergone only limited confidentiality edits as part of their disclosure avoidance. Unfortunately, record-level data are exceedingly difficult to protect in a way that offers real privacy protection while leaving the data useful for unspecified analytical purposes. At present, the Census Bureau advises research users who require such data to consider restricted-access modalities[1].

Difficulties Arising from Increased Transparency. Most users of the 2000 and 2010 Censuses were not aware of the details of the disclosure avoidance mechanism nor its impact on their results. With 2020 Census data, users will be aware that noise has been added, and they will be able to calculate the margin of error that the noise introduces. Some data users are confused about this *margin of error*, a term that they traditionally associate with sample surveys. While coverage error has long been openly discussed and analyzed,[18] discussion of the error caused by disclosure avoidance procedures, historically called "confidentiality edits," has been terse and limited to qualitative statements[20].

Misunderstandings about Randomness and Noise Infusion. A key mechanism of differential privacy is adding random noise to tabulated data before releasing. By design, the noise-injection mechanisms used by the Census Bureau will result in increased accuracy as population sizes increase. Explaining this to data users, community leaders and the general public will be critical to the acceptance of this new disclosure avoidance methodology.

For example, some statistical programmers want to use repeatable random number generators for regression testing and production, and have the ability to re-run the privacy mechanism if the first set of coins produces results that are deemed unacceptable. Differential privacy is clearly incompatible with this notion.

Although there are many technical papers explaining differential privacy, including the Harvard University Privacy Tools Project[19] and the Duke University tutorials[16], their academic language is not accessible to many of the Census Bureau's data users. The lack of simplified materials to promote a general understanding of differential privacy increases the likelihood of misunderstanding.

4 RECOMMENDATIONS

Despite the numerous challenges differential privacy adoption faced, it has taken root in the Census Bureau. Here, we present recommendations for furthering its integration into the Census Bureau and overcoming some of the hurdles outlined above.

Repeated Discussions with Decision Makers. The deployment of differential privacy within the Census Bureau marks a sea change for the way that official statistics are produced and published. But despite the problems encountered, the Census Bureau has not reconsidered its decision to adopt modern disclosure avoidance mechanisms. We believe that this is a result of the Census Bureau's longstanding commitment to confidentiality protections and the adoption of modern methodological techniques. Repeated discussions with both the Census Bureau's governing boards and with data users are vital in assembling and maintaining institutional support for this transformative effort.

Controlled Vocabulary. The Census Bureau has found it helpful to establish a controlled vocabulary of terms for discussions of matters involving differential privacy. In computing and mathematics, it is common for practitioners to adopt many different words to mean the same thing (and, conversely, to use the same words to mean different things in different contexts). Internal comprehension as well as the ease of communicating with others has been helped by having a controlled vocabulary, enforced from the highest levels of technical management.

Integrated Communications. The Census Bureau has created a communications team staffed with senior members of several directorates for the purpose of working with data users and the public on promoting understanding of the new privacy initiative. This team plays a pillar role in the acceptance of differential privacy, both internally and externally to the Census Bureau. With a public-facing educational tutorial forthcoming, and a suite of informative media in the works, they are making user-level understanding of differential privacy rapidly more available to non-experts.

Finally, the Census Bureau is expanding its educational efforts on the topic of differential privacy.

5 CONCLUSION

The Census Bureau is now two years into the process of modernizing its disclosure avoidance systems to incorporate formal privacy protection techniques. Although this process has proven to be challenging across disciplines, it promises to reward the efforts of the Census Bureau. In order to attempt privacy protection on the same scale without differential privacy, the Census Bureau could publish dramatically fewer tables and simply hope that they haven't leaked enough information to allow an attacker to perform database reconstruction. By implementing differential privacy, the Census Bureau can mathematically limit the privacy loss associated with each publication. Beyond the 2020 Census, the Census Bureau intends to use differential privacy or related formal privacy systems to protect all of its statistical publications.

It is noteworthy that this institution is not only implementing differential privacy in its statistical analyses, but truly integrating it into its organizational structure. With staff in communications, research, statistics, and computer science familiar with and supportive of differential privacy, a set of diverse employees equipped

with privacy tools will be available in the Census Bureau beyond the 2020 Census. The methods put in place for the 2018 and 2020 implementations will act as templates, greatly easing its adoption in future statistical projects. With skilled staff and effective methodology in place, differential privacy can make lasting improvements to privacy protection at the federal government's largest statistical agency.

DISCLAIMER: This paper is presented with the hope that its content may be of interest to the general statistical community. The views in these papers are those of the authors, and do not necessarily represent those of the Census Bureau.

REFERENCES

[1] 2018. Restricted-Use Microdata. https://www.census.gov/research/data/restricted_use_microdata.html#CRE1 Last Accessed July 14, 2018.

[2] John Abowd. 2016. Why Statistical Agencies Need to Take Privacy-loss Budgets Seriously, and What It Means When They Do. *Labor Dynamics Institute* (Dec. 7 2016). http://digitalcommons.ilr.cornell.edu/ldi/32/

[3] John Abowd. 2017. How Will Statistical Agencies Operate When All Data Are Private? *Journal of Privacy and Confidentiality* 7 (2017). Issue 3. https://doi.org/10.29012/jpc.v7i3.404.

[4] John M. Abowd and Ian M. Schmutte. [n. d.]. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review* ([n. d.]). https://arxiv.org/abs/1808.06303 forthcoming.

[5] Katherine G. Abraham, Ron Haskins, Sherry Glied, Robert M. Groves, Robert Hahn, Hilary Hoynes, Jeffrey B. Liebman, Bruce D. Meyer, Ron Haskins, Paul Ohm, Nancy Potok, Kathleen Rice Mosier, Robert J. Shea, Latanya Sweeney, Kenneth R. Troske, and Kim R. Wallin. 2017. *The Promise of Evidence-Based Policymaking*. Comission on Evidence-Based Policymaking, Washington, DC. https://www.cep.gov/cep-final-report.html

[6] Fredrik Andersson, John M. Abowd, Matthew Graham, Jeremy Wu, and Lars Vilhuber. 2009. Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-use Data. In *Joint NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection*. Cornell University, Suitland, MD. https://ecommons.cornell.edu/handle/1813/47672

[7] 2017. *Differential Privacy*. Apple Computer. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

[8] Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. 1974. *Discrete Multivariate Analysis: Theory and Practice*. Springer. https://www.springer.com/us/book/9780387728056

[9] Bolin Ding, Jana Kulkarni, and Sergey Yekhanin. 2017. Collecting telemetry data privately. https://www.microsoft.com/en-us/research/blog/collecting-telemetry-data-privately/

[10] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Conference on Theory of Cryptography (TCC'06)*. Springer-Verlag, Berlin, Heidelberg, 265–284. https://doi.org/10.1007/11681878_14

[11] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. In *Foundations and Trends in Theoretical Computer Science*, Vol. 9. NOW, 211–407.

[12] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. ACM, New York, NY, USA, 1054–1067. https://doi.org/10.1145/2660267.2660348

[13] FCSM 2005. *Working Paper 22: Report on Statistical Disclosure Limitation Methodology*. Technical Report. Federal Committee on Statistical Methodology. https://fcsm.sites.usa.gov/reports/policy-wp/

[14] Simson L. Garfinkel. 2018. Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance System as Implemented for the 2018 End-to-End Test. https://www.census.gov/about/cac/sac/meetings/2017-09-meeting.html

[15] Amy Lauger, Billy Wisniewski, and Laura McKenna. 2014. *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research*. Technical Report. U.S. Census Bureau. https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf

[16] Ashwin Machanavajjhala, Xi He, and Michael Hay. 2017. Differential Privacy in the Wild: A Tutorial on Current Practices & Open Challenges. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17)*. ACM, New York, NY, USA, 1727–1730. https://doi.org/10.1145/3035918.3054779

[17] Frank McSherry. 2009. Privacy Integrated Queries, In Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD). https://www.microsoft.com/en-us/research/publication/privacy-integrated-queries/

[18] Thomas Mule. 2012. *Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States*. Technical Report. U.S. Census Bureau. https://www.census.gov/coverage_measurement/pdfs/g01.pdf

[19] Kobbi Nissim, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, David O'Brien, and Salil Vadhan. 2018. Differential Privacy: A Primer for a Non-technical Audience (Preliminary Version). *Vanderbilt Journal of Entertainment and Technology Law* (2018). Forthcoming.

[20] U.S. Census Bureau 2012. *2010 Census Summary File 1: 2010 Census of Population and Housing, Technical Documentation*. Technical Report. U.S. Census Bureau. https://www.census.gov/prod/cen2010/doc/sf1.pdf

[21] US Census Bureau 2017. *Administrative Records Modeling Update for the Census Scientific Advisory Committee*. Technical Report. US Census Bureau. https://www2.census.gov/cac/sac/meetings/2017-03/admin-records-modeling.pdf

[22] U.S. Census Bureau. 2017. Our Mission. https://www.census.gov/about/what.html

EXHIBIT 2

U.S. Census Bureau, 2020 Disclosure Avoidance System Updates

https://www.census.gov/newsroom/press-releases/2020/statement-covid-19-2020.html

U.S. Department of Commerce Secretary Wilbur Ross and U.S. Census Bureau Director Steven Dillingham Statement on 2020 Census Operational Adjustments Due to COVID-19

FOR IMMEDIATE RELEASE: MONDAY, APRIL 13, 2020

APRIL 13, 2020
RELEASE NUMBER CB20-RTQ.16

APRIL 13, 2020 — The 2020 Census is underway and more households across America are responding every day. Over 70 million households have responded to date, representing over 48% of all households in America. In light of the COVID-19 outbreak, the U.S. Census Bureau is adjusting 2020 Census operations [https://2020census.gov/en/news-events/operational-adjustments-covid-19.html] in order to:

- Protect the health and safety of the American public and Census Bureau employees.
- Implement guidance from federal, state and local authorities.
- Ensure a complete and accurate count of all communities.

The Census Bureau temporarily suspended 2020 Census field data collection activities in March. Steps are already being taken to reactivate field offices beginning June 1, 2020, in preparation for the resumption of field data collection operations as quickly as possible following June 1.

In-person activities, including all interaction with the public, enumeration, office work and processing activities, will incorporate the most current guidance to promote the health and safety of staff and the public. This will include recommended personal protective equipment (PPE) and social distancing practices.

Once 2020 Census data collection is complete, the Census Bureau begins a lengthy, thorough and scientifically rigorous process to produce the apportionment counts, redistricting information and other statistical data products that help guide hundreds of billions of dollars in public and private sector spending per year.

In order to ensure the completeness and accuracy of the 2020 Census, the Census Bureau is seeking statutory relief from Congress of 120 additional calendar days to deliver final apportionment counts.

Under this plan, the Census Bureau would extend the window for field data collection and self-response to October 31, 2020, which will allow for apportionment counts to be delivered to the President by April 30, 2021, and redistricting data to be delivered to the states no later than July 31, 2021.

###

Contact

Public Information Office
301-763-3030
pio@census.gov [mailto:pio@census.gov]

Related Information

2020 Census Operational Adjustments Due to COVID-19 [https://2020census.gov/en/news-events/operational-adjustments-covid-19.html]

EXHIBIT 3

U.S. Department of Commerce Secretary Wilbur Ross and U.S. Census Bureau Director Steven Dillingham Statement on 2020 Census Operational Adjustments Due to COVID-19 U.S. Census Bureau (Apr. 13, 2020)

https://www.census.gov/newsroom/press-releases/2020/statement-covid-19-2020.html

U.S. Department of Commerce Secretary Wilbur Ross and U.S. Census Bureau Director Steven Dillingham Statement on 2020 Census Operational Adjustments Due to COVID-19

FOR IMMEDIATE RELEASE: MONDAY, APRIL 13, 2020

APRIL 13, 2020
RELEASE NUMBER CB20-RTQ.16

APRIL 13, 2020 — The 2020 Census is underway and more households across America are responding every day. Over 70 million households have responded to date, representing over 48% of all households in America. In light of the COVID-19 outbreak, the U.S. Census Bureau is adjusting 2020 Census operations [https://2020census.gov/en/news-events/operational-adjustments-covid-19.html] in order to:

- Protect the health and safety of the American public and Census Bureau employees.
- Implement guidance from federal, state and local authorities.
- Ensure a complete and accurate count of all communities.

The Census Bureau temporarily suspended 2020 Census field data collection activities in March. Steps are already being taken to reactivate field offices beginning June 1, 2020, in preparation for the resumption of field data collection operations as quickly as possible following June 1.

In-person activities, including all interaction with the public, enumeration, office work and processing activities, will incorporate the most current guidance to promote the health and safety of staff and the public. This will include recommended personal protective equipment (PPE) and social distancing practices.

Once 2020 Census data collection is complete, the Census Bureau begins a lengthy, thorough and scientifically rigorous process to produce the apportionment counts, redistricting information and other statistical data products that help guide hundreds of billions of dollars in public and private sector spending per year.

In order to ensure the completeness and accuracy of the 2020 Census, the Census Bureau is seeking statutory relief from Congress of 120 additional calendar days to deliver final apportionment counts.

Under this plan, the Census Bureau would extend the window for field data collection and self-response to October 31, 2020, which will allow for apportionment counts to be delivered to the President by April 30, 2021, and redistricting data to be delivered to the states no later than July 31, 2021.

###

Contact

Public Information Office
301-763-3030
pio@census.gov [mailto:pio@census.gov]

Related Information

2020 Census Operational Adjustments Due to COVID-19 [https://2020census.gov/en/news-events/operational-adjustments-covid-19.html]

EXHIBIT 4

2020 Census Response Rate Update: 99.98% Complete Nationwide, U.S. CENSUS BUREAU (Oct. 19, 2020)

https://www.census.gov/newsroom/press-releases/2020/2020-census-all-states-top-99-percent.html#:~:text=OCT.,15%2C%202020.

2020 Census Response Rate Update: 99.98% Complete Nationwide

FOR IMMEDIATE RELEASE: MONDAY, OCTOBER 19, 2020

OCTOBER 19, 2020
RELEASE NUMBER CB20-CN.120

All States Top 99%, Self-Response Beats 2010

OCT. 19, 2020 — According to updated numbers released by the U.S. Census Bureau today, 99.98% of all housing units and addresses nationwide were accounted for in the 2020 Census as of the end of self-response and field data collection operations on Oct. 15, 2020. In all states, the District of Columbia and the Commonwealth of Puerto Rico, more than 99% of all addresses have been accounted for, and in all but one state that number tops 99.9%.

“The 2020 Census faced challenges like no other decennial census in living memory,” said Secretary of Commerce Wilbur L. Ross, Jr. “Achieving these metrics in the face of severe weather events and a global pandemic is a testament to the determination and ingenuity of the hundreds of thousands of dedicated women and men who worked on the 2020 Census.”

Compared to the final self-response rate of 66.5% for the 2010 Census, 67% were accounted for through self-response to date, with the rest having been accounted for through our Nonresponse Followup (NRFU) operation.

“America stepped up and answered the call: shape your future by responding to the 2020 Census,” said Dr. Steven Dillingham, Director of the Census Bureau. “Generally, better data comes from self-response, but after a decade of global decline in census and survey participation along with the challenges presented to communities by COVID-19, we had not expected to exceed the 2010 self-response rate. That we did is a testament to the American people, our nearly 400,000 national and community partners, and very importantly our staff.”

“The Census Bureau was able to meet and overcome many challenges because of our innovative design and use of new technology, but it could not be done without the unflinching resolve of our staff,” Dillingham continued. “We thank everyone on the team for their contributions, from the census takers and field staff going the extra mile to reach those hardest to count, to the dedicated operational leadership at headquarters and around the country working around the clock to maintain and protect our systems, process the data, oversee the operation, and get the word out about the importance of the 2020 Census.”

“We are especially proud of the hard work done to bring the state of Louisiana over 99% complete despite the devastating effects of hurricanes Laura and Delta, and of the partnership with American Indian and Alaska Native tribal governments to get 99.77% of the NRFU workload on their lands done, despite closures due to the pandemic.”

“Hundreds of millions of people were counted in the 2020 Census, and statisticians and data quality experts are now busy making sure everyone was counted once, only once, and in the right place,” Dillingham continued. “The Census Bureau will use the best methodologies available to resolve the very small number of unresolved addresses and to ensure that our data products are accurate.”

Each census, the Census Bureau produces coverage estimates [/programs-surveys/decennial-census/about/coverage-measurement.html] and conduct extensive assessments that we share with the public. The completion rates are just early indicators. For more information on the 2020 Census, including use of proxy and administrative records, please see our updated FAQ [/newsroom/press-kits/2020/2020-census-faqs.html] s.

The Census Bureau is working hard to process the data in order to deliver complete and accurate state population counts as close as possible to the Dec. 31, 2020, statutory deadline.

Data collection for the 2020 Census ended at 11:59 p.m. Hawaii Standard Time on Oct. 15, 2020 (5:59 a.m. EDT). Paper responses are still arriving and will be processed if postmarked by October 15, and received at the processing center no later than October 22.

For more information, visit 2020census.gov [https://2020census.gov/] .

###

Contact

Public Information Office
301-763-3030
pio@census.gov [mailto:pio@census.gov]

Related Information

2020 Census FAQs
Press kit [https://www.census.gov/newsroom/press-kits/2020/2020-census-faqs.html]

EXHIBIT 5

Albert E. Fontenot, 2020 Census Update, Presentation to the Census Scientific Advisory Committee March 18, 2021 at 2

https://www2.census.gov/about/partners/cac/sac/meetings/2021-03/presentation-2020-census-operational-review.pdf

2020 Census Update

Presentation to the Census Scientific Advisory Committee
March 18, 2021

Albert E. Fontenot, Jr., Associate Director Decennial Census Programs

Deborah M. Stempowski, Assistant Director for Decennial Census Programs, Operations and Schedule Management

Operational Timelines: Original and Pandemic-Adjusted

Activity / Operation	Original Dates	Replan Dates (as presented to CSAC on Sept. 17, 2020)	Final Dates
Update Leave (Stateside)	March 15 – April 17	Phased re-opening occurred between May 4 and June 12	Phased re-opening occurred between May 4 and June 12
Service Based Enumeration	March 30 – April 1	September 22 – 24	September 22 – 24
Targeted Non-Sheltered Outdoor Locations	March 31 – April 1	September 23 – 24	September 23 – 24
Group Quarters Enumeration	April 2 – June 5	April 2 – September 3	April 2 – September 3
Enumeration of Transitory Locations	April 9 – May 4	September 3 – 28	September 3 – 28
Nonresponse Followup*	**May 13 – July 31**	**August 9 – September 30**	**August 9 – October 15**
Delivery of Apportionment Data**	**By Statutory Deadline: December 31, 2020**	**By Statutory Deadline: December 31, 2020**	**April 30, 2021**
Delivery Redistricting Data**	**By Statutory Deadline: March 30, 2021**	**Plan in Development**	**September 30, 2021**

*For a period of time, NRFU was 8/11/20-10/31/20.

**For a period of time, delivery of apportionment data by 4/30/21 and redistricting data by 7/31/21, were considered.

2020 Census
Data Products: Redistricting Legacy Format Summary File

We previously announced that we would deliver redistricting data to the states and the public by September 30, 2021. This was based on our thorough examination of the revised post processing schedule, and our focus on fulfilling our constitutional obligation to deliver the state population counts for apportionment to the President.

This is problematic for some states, and as such, we have been and continue to explore alternatives to provide this data to the states as quickly as possible.

One alternative is delivery of a Legacy Format Summary File, which could be delivered by mid-to-late August 2021. This data set:

- Provides a possible source for states with a pressing need to access redistricting data
- Fully reviewed and cleared for publication by mid-to-late August 2021
- Uses the same data as what will be released in more user-friendly format by September 30, 2021
- Product was always part of the 2020 Census product plan
- Format produced and provided to the states since at least Census 2000
- Prototype data in this format from the 2018 End-to-End Census Test is available for designing and testing redistricting systems
- Requires additional handling to properly extract data from this format

2020 Census
Summary of Self-Response

Original Dates: March 12 – July 31, 2020

Adjusted Dates: March 12 – October 15, 2020

- **Final Self-Response Rate: 67.0%**
  - Exceeded Final 2010 Census Self-Response Rate of 66.5%
- Self-Response Volumes by Mode:
  - **Total: 99.02 million self-responses**
  - **Internet: 79.08 million (79.86%)**
  - **Paper: 18.11 million (18.29%)**
  - **Phone: 1.83 million (1.85%)**

- 14 States with a Self-Response Rate at or above 70% vs 7 States in 2010
- 47 States with a Self-Response Rate at or above 60% vs 47 States in 2010
- 28 States that met or exceeded their final 2010 Census Self-Response Rate

2020 Census
Nonresponse Followup Summary

- Operational Dates: August 9 – October 15, 2020
- Successful implementation of a rolling soft launch that began July 16, 2020
- **Completed Housing Units (HUs): 60.8M**
  - **Completed via Self-Response: 6.3M** (these are included in the total self-response rate of 67%)
  - **Total Enumerated Occupied HUs: 30.7M**
    - **Enumerated via Householder: 17.1M (55.6%)**
    - **Enumerated via Proxy: 7.4M (24.1%)**

      This proxy response rate of 24.1% is similar to the 2010 proxy response rate of 23.8%.
    - **Enumerated via Administrative Records: 6.3M (20.4%)***
  - **Vacant HUs: 13.5M**
  - **Deleted HUs: 10.3M**

Approximately 13.9% of the full NRFU workload (including vacant and deleted housing units) were completed using high-quality administrative records, lower than the expected rate of 22.5%.

Note: All numbers are subject to change upon completion of post collection processing.

2020 Census
Overall Data Collection Successes

- **99.9% resolution** In all 50 states, the District of Columbia and the Commonwealth of Puerto Rico, more than 99% of all addresses have been resolved. In all but one state that number tops 99.9%.

- **2 in 3 households responded on their own**
  - Final self-response rate of 67.0%, exceeding the final self-response rate of 66.5% for the 2010 Census.
  - 99.0M Self-Responding Housing Units (79.8% responded by internet, 18.3% by paper, 1.9% by phone)

- **Not 1 second of downtime on ISR** Internet Self-Response option successfully managed our highest traffic demand and operated throughout the census without one second of downtime.

- **Increased use of technology at every level** Automation and increased use of technology such as enumerator use of iPhones for case routing optimization, assignment management, and data collection contributed to increased enumerator productivity.

- **1.92 cases completed per hour** Achieved enumerator productivity rate of 1.92 cases per hour, compared to 1.05 cases per hour for the 2010 Census.

2020 Census
Factors that Enabled Progress

- Smooth Launch of Self-Response options - Online, by Phone, by Paper
- Phased resumption of field data collection activities
- Transitioned key training activities from classroom to virtual training, affecting training for nearly 500,000 workers
- Incorporated the use of pay flexibilities to minimize turnover of trained operational staff
- Used alternate means of data collection, including adapted processes to incorporate broader use of administrative records, such as lists of students from colleges and universities
- Remained flexible and agile to adapt to ever changing on-the-ground conditions, including instituting outbound telephone enumeration and additional mailings.
- Contingency funding, on various fronts, supported operational adjustments necessary to complete data collection.

2020 Census
Post Processing

Post processing activities are conducted once all data collection is complete

- Turning all of the response data we received into usable statistics is complex work that is guided by our statistical quality standards.
- We start by taking all of the responses we received across response modes and operations and integrate that data with our information about addresses. We then follow established statistical methods for verifying whether we have a response from every address, resolving duplicates, and filling in missing information.
- Just as we did during data collection, we are continuously checking the quality of the data throughout data processing.
- As with all prior censuses we have found issues as we prepare the data for tabulation. While some issues appear to be pandemic related, most are what we experience with every decennial census and other Census Bureau surveys. We expect these kinds of anomalies and issues, and they are similar to the Census Bureau's experience in prior decennial censuses.
- Importantly, we have not uncovered anything so far that would suggest that the 2020 Census will not be fit for its constitutional and statutory purposes.
- **Census Bureau is working to thoroughly correct and address all issues and anomalies** as a part of our mission to deliver accurate 2020 Census data products as close to the statutory deadline as possible.

2020 Census
Data Products: First Tier

Apportionment Product – by April 30, 2021
The Apportionment Product will be the first release of the 2020 Census and will provide the apportionment population and the number of seats in the U.S. House of Representatives by state. The product will also include the resident population of the 50 states plus the overseas federal employees (military and civilian) and their dependents living with them, who are included in their home states.

Redistricting File (P.L. 94-171) – by September 30, 2021
Public Law 94-171 directs the Census Bureau to provide data to the governors and legislative leadership in each of the 50 states for redistricting purposes. This product will be the first file released that will include demographic and housing characteristics about detailed geographic areas.

Demographic Profile
This product will provide critical demographic and housing characteristics about local communities as soon after the release of the Redistricting file as possible.

Demographic and Housing Characteristics File (DHC)
The DHC will include many of the demographic and housing tables previously included in Summary File 1.

2020 Census

Ensuring High Quality Data from the 2020 Census

The 2020 Data Quality Executive Governance Group (EGG) was chartered in April 2020 by the Deputy Director/Chief Operating Officer to ensure that we had the right focus and resources dedicated to detecting and addressing data quality issues related to the 2020 Census. The EGG is comprised of career technical leadership and led by the Associate Director for Demographic Programs and Chief Demographer, Associate Director for Research and Methodology and Chief Scientist, and Assistant Director for Decennial Programs, Operations and Schedule Management. This new special team, with expertise from the entire Census Bureau, supplements the existing expert teams and provides extra focus on data quality.

Working Groups	Objectives	Deliverables
Lead Operational Update Team **Administrative Records Usage Team** **Demographic and Housing Reasonableness Review "CUF/CEF"** **Demographic Analysis and Population Estimates** **Post Enumeration Survey** **Current Surveys Field Experience Team***	**Existing Teams**: • Continue current work **New + Existing Teams**: • Identify new/emerging ways to assess and/or ensure quality (real time and post-data collection)	Operational changes and data quality assessments will be documented by the **Data Quality Documentation Team***

*New team, not previously part of 2020 Census operations

2020 Census
Data Quality Assessment Efforts and Timeline

Asking outside experts to review our work is standard operating procedure at the U.S. Census Bureau. It underscores our commitment to quality and transparency.

- **Releasing** information and metrics on data quality on an earlier schedule than typical with a decennial census.
- **Leveraging** external engagement opportunities with organizations such as the American Statistical Association.
- **Engaging** with the National Academy of Sciences (NAS) Committee on National Statistics, American Statistical Association Quality Indicators Task Force, and JASON.
    - These three groups will tackle different aspects of assessing the Census Bureau's work. Their reports will advise the Census Bureau on improving future censuses and will help the public understand the quality of the 2020 Census data.
    - JASON report, *Assessment of 2020 Census Data Quality Processes*, was released on February 23, 2021.
- **Exploring** additional quality assessments, beyond those planned in operational assessments and evaluations.

Critical Milestones for Release of Data Quality Assessment Metrics:

- Release of Demographic Analysis Results: Released December 15, 2020
- Release of 2020 Census Operational Quality Metrics to accompany Resident Population Counts: April 2021
- Release of 2020 Census Operational Quality Metrics to accompany Redistricting Data Products: September 2021

2020 Census

Upcoming 2020 Census Research Publications: Assessments and Evaluations

Assessments are designed to document final volumes, rates, and costs for individual operations or processes using data from production files and activities and information collected from debriefings and lessons learned. A total of 54 Operational Assessments on the 2020 Census will be published, beginning in Summer 2021. Assessments of note include:

- In-Office Address Canvassing Operational Assessment (Summer 2021)
- In-Field Address Canvassing Operational Assessment (Summer 2021)
- Demographic Analysis Operational Assessment (Summer 2022)
- Nonresponse Followup Operational Assessment (Fall 2022)

Evaluations are designed to analyze, interpret, and synthesize the effectiveness and efficiencies of census components and their impact on data quality and coverage. A total of 14 Evaluations on the 2020 Census will be published, beginning in Spring 2021. Evaluations of note include:

- Research on Hard to Count Populations: Non-English Speakers and Complex Household Residents including Undercount of Children (Spring 2022)
- Analysis of Census Internet Self-Response Paradata by Language (Winter 2022)

2020 Census
Communications and Blog Post Plan

What's Planned
- A series of accessible blogs in the voices of the Census Bureau's internal, career experts about the quality and progress of the 2020 Census.
- We published a series of similar blogs prior to release of the 2010 apportionment counts.
- One or two blogs per week, corresponding to other releases and events.

Why We're Doing It
- To help restore the Census Bureau's credibility as an independent statistical agency
- To educate the public and our stakeholders about the quality of the 2020 Census
- To set expectations for what is coming

Overall Status
- Seven blogs posted already.
- Fluid schedule – we may add additional subjects.

2020 Census
Blog Post Schedule

Blog Title	Tentative Date
Pandemic and All its Effects	Feb. 2, 2021 **POSTED**
Census Processing 101	Feb. 11, 2021 **POSTED**
Timeline Context for Redistricting	Feb. 12, 2021 **POSTED**
Ensuring a Robust and Accurate Data Quality Analysis in the 2020 Census	Feb. 23, 2021 **POSTED**
Adapting Field Operations to Meet Unprecedented Challenges	March 1, 2021 **POSTED**
Finding 'Anomalies' Illustrates 2020 Census Quality Checks Are Working	March 9, 2021 **POSTED**
2020 Census Group Quarters	March 16, 2021 **POSTED**
Introduction to Quality Indications	Mid March
Administrative Records	Late March
Subject Matter Expert Review Process	Late March
Imputation	Late March
Unduplication	Late March
Post Enumeration Survey	Early April
Apportionment Process and What to Expect on Release Day	Early April
Director's Blog about first 2020 Census Data Release	April 16-30, 2021
Comparisons to Benchmarks and Examining Operational Quality Metrics	April 16-30, 2021
Release of Table 2 of quality metrics	Mid to late May

Thank You

Albert E. Fontenot, Jr.

Associate Director for Decennial Census Programs

Deborah M. Stempowski

Assistant Director for Decennial Census Programs, Operations and Schedule Management

U.S. Department of Commerce
U.S. Census Bureau
4600 Silver Hill Rd.
Suitland, Maryland 20746

EXHIBIT 6

Michael B. Hawes, U.S. Census Bureau, Implementing Differential Privacy: Seven Lessons From the 2020 United States Census, Harvard Data Science Review (Apr. 30, 2020)

https://hdsr.mitpress.mit.edu/pub/dgg03vo6/release/2

Issue 2.2 ••• Published on Apr 30, 2020 DOI 10.1162/99608f92.353c6f99 SHOW DETAILS

Implementing Differential Privacy: Seven Lessons From the 2020 United States Census

by Michael B. Hawes

Published on Apr 30, 2020

last released 7 months ago

ABSTRACT

With the 2020 Census now underway, there is substantial national and global interest in the U.S. Census Bureau's decision to modernize the statistical safeguards that will be used to protect respondent privacy. The Census Bureau's adoption of differential privacy
Census marks a transformation
official statistics. But, the transi
differential privacy has raised a
questions about the proper bala

privacy and accuracy in official statistics, the prioritization of certain data uses over others, and the future of statistical offices and their data products. As organizations increasingly consider differential privacy as a solution to the vexing privacy threats of today, the Census Bureau's experiences in navigating these issues may be instructive for statistical agencies, corporations, researchers, and data users across the United States, and around the world.

Keywords: census, differential privacy, disclosure avoidance, statistical infrastructure, official statistics, data privacy

1. The 2020 Census Is Underway

April 1, 2020, is Census Day in the United States. Mandated by Article I, Section 2 of the U.S. Constitution, this once-every-decade data collection is the nation's largest civilian mobilization effort. The 2020 Census[1] is projected to cost approximately $15.6 billion from start to finish. While the majority of households across the nation are expected to self-respond (by internet, teleph
the U.S. Census Bureau is expec
500,000 temporary census take
door-to-door to those household
respond, in an effort to count ev

every household "*once, only once, and in the right place*." Field operations are critical to the success of this operation, both to process the millions of paper questionnaires at the Census Bureau's processing centers, and to visit those households that do not self-respond. However, "In light of the COVID-19 outbreak, the U.S. Census Bureau has adjusted 2020 Census operations" in order to "protect the health and safety of Census employees and the American public" and to "ensure a complete and accurate count of all communities" (U.S. Census Bureau, 2020b). To that end, "The 2020 Census is adaptable and equipped with an approximate $2 billion dollar contingency budget for circumstances like the COVID-19 outbreak" (U.S. Department of Commerce, 2020).

The expense and logistics involved in this recurring enumeration of the nation's population may seem daunting, but the data collected are critical to decision-making at all levels of society. Census data are used to apportion seats in the U.S. House of Representatives, help guide the allocation of approximately $675 billion in federal funds each year based on population counts and demographic characteristics (Hotchkiss & Phelan, 2017), support critical private sector decision-making at the national, state, and local levels, and serve as the benchmark

statistics for incumbent elected officials and policymakers throughout the subsequent decade (Sullivan, 2020). Throughout its history, the Census Bureau has taken this responsibility seriously, and with each census, the Census Bureau improves and adapts its methods in an effort to produce the most complete and accurate count possible.

Producing accurate data to support these important functions is central to the Census Bureau's mission "to serve as the nation's leading provider of quality data about its people and economy." But, in fulfilling this responsibility, the Census Bureau is also required to ensure the privacy of its respondents and the confidentiality of their data. Title 13, Section 9 of the U.S. Code[2] prohibits the agency from disclosing any personally identifiable information in its statistics and data products. To balance these countervailing responsibilities, the Census Bureau has long been a world leader in the design and innovation of statistical methods that minimize the likelihood that individuals can be reidentified in its public data products. As statistical offices and other data-centric organizations around the world know, however, recent advances in computing power and the proliferation of new data sources make this task increasingly difficult. Recent internal experiments at the Census

Bureau sought to assess this growing vulnerability, and the results were alarming. Using only a portion of the publicly released aggregate data from the 2010 Census,[3] Census Bureau researchers were able to accurately reconstruct individual-level records with selected attributes for the entire U.S. population, and were able to accurately determine location (census block), age (+/- one year), sex, race (63 categories), and ethnicity for 219 million individuals. At that degree of precision, more than 50% of all persons censused in 2010 were unique within the population. Matching these individual records against commercially available data from 2010 to attach names to these records, the Census Bureau was able to confirm accurate reidentifications for 52 million people (Abowd, 2019). Recognizing that the traditional statistical techniques used to protect privacy in prior decades are increasingly insufficient to counter the privacy threats of today, the Census Bureau decided to modernize its approach to data protection and has committed to using *differential privacy* to protect the confidentiality of the 2020 Census (U.S. Census Bureau, 2019).

This journal has previously exa
of differential privacy, includin
discussion by Daniel L. Oberski
Kreuter in volume 2.1 (2020), so I would direct

readers interested in learning more about differential privacy as an approach to those articles. Instead, I would like to highlight some of the lessons that the Census Bureau has learned so far from its implementation of differential privacy. Statistical offices, corporations, researchers, and data users across the United States and around the world are watching the Census Bureau's adoption of differential privacy with keen interest, and I hope that our experiences (and miss-steps) along the way will prove instructive.[4]

2. Seven Lessons Learned From the Implementation of Differential Privacy for the 2020 Census

2.1. Lesson One: The Emerging Public Policy Debate About Privacy and Accuracy

The database reconstruction theorem, also known as the fundamental law of information reconstruction, tells us that if you publish too many statistics derived from a c source, at too high a degree of a after a finite number of queries completely expose the confiden & Nissim, 2003). All statistical d limitation techniques, including traditional and

limitation techniques, including traditional and formally private methods, seek to protect privacy by limiting the quantity of data released (e.g., through suppression) or by reducing the accuracy of the data. It should be noted that the impacts of any privacy protection method on data availability or accuracy should not be seen as technical byproducts; protecting respondent privacy fundamentally requires reducing one or both of these dimensions to be effective. Protection methodologies that rely on suppression or coarsening of the data can have significant impacts on data usability, but these methods have generally been tolerated by data users because the reasons for their use are fairly intuitive; the link between small cell counts or highly precise statistics and the identities of specific individuals is easy to grasp. Methodologies that rely on noise injection to protect privacy also have not received much criticism by data users largely because these methods' impact on data accuracy are not typically observable; in most implementations the swapping rates or noise injection parameters and assessments about their impact on accuracy are kept confidential to prevent reverse engineering of the original, con

The transparency of formally pr
and the explicit quantification c
accuracy tradeoff through the u
loss budgets (epsilon), have enabled data users

and privacy advocates, however, have voiced

debate the relative importance of accuracy vs. privacy in unprecedented ways. Where policy decisions about what constituted "sufficient" protection or accuracy were previously made behind closed doors, differential privacy has brought that debate into the court of public opinion. Over the last few months, the Census Bureau has learned the hard way that navigating this debate is difficult, but essential, to maintaining both the public's trust in the proper safeguarding of their information, and the credibility of the data products on which data users rely.

2.2. Lesson Two: Prioritizing Accuracy for Diverse Use Cases

When implementing differential privacy, the privacy-loss budget makes data accuracy and privacy competing uses of a finite resource: the information (bits) in the underlying data (Abowd & Schmutte, 2019). It is impossible to protect privacy while also releasing highly accurate data to support every conceivable use case, and vice versa. While statistics for large populations—for example, for e
for major metropolitan areas—c
adequately protected with negli
noise, many important uses of c
require calculations on smaller

where the impacts of noise can be much more significant. When designing the differentially private systems that will be used for the 2020 Census, the Census Bureau had to start by enumerating the myriad ways that census data are used and identifying which of those uses are more critical than others.[5] Some priority use cases are obvious: those that support congressional and state legislative redistricting, for example, or those that enable the equitable and efficient allocation of federal or state funding. But, deciding the relative priority of other important uses of census data is more difficult. For example, should more of the privacy-loss budget be expended on statistics that allow municipalities to know where to build hospitals and schools, or should it be spent on benchmark statistics that serve as the sampling frame and survey weights for demographic and health care surveys throughout the decade?

The relative prioritization of these use cases among many others, and the implications that they have on the design and implementation of a differentially private system cannot be made without extensive engagement and discussion with the various data user comm making these decisions, the Cen had to rely on the expert advice advisory committees, formal co American Indian and Alaska Native tribal

leaders, ongoing engagement with state and local governments, data user groups, and professional associations, as well as feedback from the public at large. While statistical offices and their data users may find this type of engagement challenging—pitting the relative importance of accuracy for one group of data users over another—it is worth considering that having these debates represents an improvement over the *status quo ante*. Most uses of traditional disclosure avoidance methods, like data swapping, required making similar tradeoffs, but the confidential nature of those methods' design, parameters, and impacts on accuracy (considered necessary to prevent reverse engineering of the confidential data) meant that these were internal agency decisions. The increased transparency that differential privacy permits now allows these trade-offs, and their consequences, to be publicly discussed and debated.

2.3. Lesson Three: Choose the Right Design

How you implement any disclosure avoidance strategy will impact the accurac
of the resulting data, and this is
for differentially private metho
design of the system can often h
impact on the accuracy of the re

than the selection of the privacy-loss budget.

With differential privacy, the amount of noise you must inject into the data is dependent on the sensitivity of the calculation you are performing. Because that sensitivity depends on the impact that the presence or absence of any individual could have on the resulting calculation, some statistics (e.g., simple counts of individuals) typically require less noise than others (e.g., mean age). But even if you are limiting your calculations to simple counting queries, the way you combine possible values of the attributes you are counting can quickly increase the sensitivity of the calculation dramatically. Take statistics about racial demographics, for example. The decennial census produces a number of tables that disaggregate population statistics by 63 values for race, that is, the six racial categories, including some other race, alone, or in any combination except "none of the above." But the census also allows individuals to write in detailed racial groups within those categories (e.g., "Scottish" or "Cherokee"), and produces other data products that disaggregate by all the permutations of those detailed g
the range of possible values for
different sets of tabulations diff
when measured alone or in com
the sensitivity of those calculati

Put another way, the mathematical framework of differential privacy shows us something that was often obscured when using traditional disclosure avoidance methods: protecting privacy is significantly more costly for some queries than for others. Adopting a one-size-fits-all approach to algorithm design for these different sets of tabulations would quickly exhaust the overall privacy-loss budget, resulting in poor accuracy across the board. Instead, to address the added sensitivity of the detailed race groupings, the Census Bureau chose to implement two different disclosure avoidance solutions for these groups of data products, each based on differential privacy and sharing the same global privacy-loss budget, but with separate algorithms designed to optimize for accuracy in different ways. Understanding the varying sensitivity of your desired statistics, and more importantly, knowing which data use cases are most important, are critical to designing the system or systems that will best meet the needs of your data users.

2.4. Lesson Four: The Bes
Plans...

Sometimes the algorithms you d
implement differential privacy
unexpected ways. The Census B

this lesson the hard way in October 2019 when it produced a set of demonstration data products that ran 2010 Census data through an early version of the Disclosure Avoidance System (DAS). In that instance, the postprocessing that the algorithm performed on the data to render it into the format traditionally associated with census results (nonnegative integers with tabular consistency) introduced far more error into the resulting data than came from the differentially private noise used to protect privacy. Even more concerning was the fact that while the differentially private noise was statistically unbiased, the postprocessing errors introduced some significant biases into the data, effectively moving people from urban centers to rural areas, among other distortions. This experience illustrates the importance of not relying on intuitive solutions without fully understanding their theoretical properties or implications; it is critical that you test and retest how that system operates in practice. Ideally, get your data users involved in the process, so that they can help you identify where and how your algorithm may not be behaving as intended. Based on the feedback received
data user groups about the dem
(Committee on National Statisti
Census Bureau is already imple
number of design changes to th

those unanticipated distortions. These efforts will continue throughout the remainder of 2020, and the Census Bureau is working closely with the data user community throughout this process (Abowd & Velkoff, 2020).

2.5. Lesson Five: Rethinking Tabular Consistency and Integrality

Consumers of official statistics, particularly those who use data products that have been produced for a long time, are accustomed to the data looking a certain way, and to interpreting those data as the 'ground truth.' As such, they are unaccustomed to seeing population counts with fractional or negative values. Because differential privacy injects noise from a symmetric distribution (typically Laplace or geometric), the raw noisy statistics emerging from the privacy protection stage of a formally private algorithm will usually include fractional and negative values, and different tabulations of the same characteristic may not be internally consistent (e.g., the total number of people in a geography may not equal the sum of males and females within that geography). The process of converting these noisy values into nonnegative integers with tabular consistenc introduces more error into the d strictly necessary to protect priv can also improve the error in so (e.g., by constraining the cumul

lower geographic levels using the counts

determined for larger geographies). As the use of differential privacy for official statistics expands, it would be advantageous for statistical agencies and data users alike to reevaluate their adherence to traditional expectations for how official statistics should look. It may be confusing to say that a town has a negative, fractional number of individuals with a particular combination of uncommon attributes, but relaxing the assumptions of nonnegativity and integrality can provide consumers of official statistics with more accurate data on a cumulative scale. Adopting these changes would require explanation and guidance on how to properly interpret these statistics, but would enable data users to effectively model the unbiased noise from the privacy protections into their analyses, improving the overall accuracy of their results. For example, by knowing the sensitivity of the query that produced a differentially private statistic, and the share of the privacy-loss budget allocated to that query, data users can calculate the probability distribution of the noise used to protect the statistic. Then, using likelihood
methods, they can factor those
probabilities into their analyses
statistical results (Abowd & Sch
Technical Appendix).

2.6. Lesson Six: Explore Alternatives

One of the largest vulnerabilities for the census--and for official statistics more broadly--is that the data are used for so many diverse purposes that supporting those uses has traditionally required publishing data at very fine levels of granularity. It is worth considering that many of these uses could be supported through alternative statistical products without the public release of the finely disaggregated data. Alternative data products could permit statistical agencies to produce more accurate inputs to these uses at a lower overall privacy risk. Tiered access models, for example, where approved data users could access the confidential data for their analysis, with noise injected to their results, have long been seen as a viable alternative for some uses. With the passage of the Foundations for Evidence-Based Policymaking Act of 2018, the federal government is exploring how to increase the availability of confidential data through tiered access methods (Potok, 2019). Synthetic data sets with validation servers, and formally private public-facing analysis engines that run calculations on the confidential
noisy results could also reduce p
for the more privacy-challengin
products on which data users ha
relied.

Another promising trend to address this issue is the blending of official tabulated statistics with official estimates based on statistical models. In many cases the other sources of uncertainty present in low-level data (operational error, coverage error, measurement error, etc.) are significant enough that data quality is already low. Statistical agencies, including the Census Bureau, release data at very detailed levels of geography to support the use case of building aggregates not elsewhere defined. It is well-known that these detailed geographical units should not be considered error-free estimates of the small area. The use of statistical models that can account for these sources of error, and that can incorporate formally private noise into the models, may yield more useful small-area data for data users than they currently get. Though it has not yet transitioned to differential privacy, the Census Bureau's Small Area Income and Poverty Estimates program, which models data from the American Community Survey, is a good example of the value of small area modeling for official statistics. As statistical offices evaluate the usefulness of their existing
especially as they consider the r
data products, it would be advar
about how they can leverage the
approaches and technologies to

useful data for their data users, without incurring the privacy risks associated with the public release of finely disaggregated data.

2.7. Lesson Seven: Remember Why We're Doing This

When engaging in discussions about the growing privacy risks, it is easy to think that the solution is to just decrease how much data you release, or to increase the privacy protections. Differential privacy certainly provides a mechanism to do this: just set your privacy-loss budget lower to compensate for the added risk. Statistical officials should, however, be wary of increasing the protections as a long-term solution. Yes, the Census Bureau, like statistical offices around the world, has a legal and ethical obligation to protect confidentiality, but the fundamental reason for operating is the production of statistical data products that support our respective societies. If the data products agencies produce lose their utility because they have lowered accuracy too much in the service of those confidentiality protections, then they should ask themselves why they are producing statistics in the first p viable and valuable to our societ agencies and the policymakers v need to consider responses to th privacy threats as part of a broa

about what official statistics are, what form they should take, what legal privacy protections they should have, and how statistical agencies can support data users in new and innovative ways.

3. How You Can Contribute

The Census Bureau will release the first of the differentially private data products from the 2020 Census in March 2021. Between now and then, there are a number of important tasks remaining to accomplish. The algorithms that will apply differential privacy on these data are already in place, but much can still be done to improve their operation and to optimize the systems to improve accuracy for the priority data use cases. Similarly, the Census Bureau must still make the final policy decisions regarding the global privacy-loss budget for the 2020 Census, as well as the final allocation of that privacy-loss budget across the various data products, tables, geographic levels, and queries. The Census Bureau will provide regular updates on these efforts via the Disclosure Avoidance and the 2020 Census webpage (https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosur 2020-census.html). Continued i data science community will be these design improvements and discussions. Readers with suggestions,

recommendations, and technical or policy considerations relating to any of the topics discussed here can submit them to the Census Bureau at 2020DAS@census.gov. The implementation of differential privacy for the 2020 Census marks a transformational moment for official statistics and for the broader data science community. Your engagement and input can help ensure that this effort will be a success.

Disclosure Statement

The views stated in this article are those of the author and not the U.S. Census Bureau. The statistics in this article have been cleared for public dissemination by the Census Bureau Disclosure Review Board (CBDRB-FY20-100).

Acknowledgments

The author drafted this article as part of his official duties as an employee of the U.S. Census Bureau. This article benefited substantially from the helpful comments and suggestions of John Abowd, Victoria Velkoff, Cynthi
Nancy Potok, Frauke Kreuter, X
Shelly Martinez, Danah Boyd, a
anonymous reviewer. The autho
to thank his many colleagues at

Bureau who contributed to the information contained herein.

References

Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* 2867. https://doi.org/10.1145/3219819.3226070

Abowd, J. M. (2019, February 16). *Staring-down the database reconstruction theorem* [Conference session]. American Association for the Advancement of Science Annual Meeting, Washington, DC. Retrieved from https://www2.census.gov/programs-surveys/decennial/2020/resources/presentations-publications/2019-02-16-abowd-db-reconstruction.pdf?.

Abowd, J. M., & Schmutte, I. M. (2016). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity, 2015*(1), 221–293. https://doi.org/10.1353/eca.2016

Abowd, J. M., & Schmutte, I. M. economic analysis of privacy pr statistical accuracy as social cho

Economic Review, *109*(1), 171–202. https://doi.org/10.1257/aer.20170627

Abowd, J., & Velkoff, V. (2020, February 12). *Census Bureau works with data users to protect 2020 Census data products*. *Research Matters Blog, U.S. Census Bureau*. https://www.census.gov/newsroom/blogs/research-matters/2020/02/census_bureau_works.html

Committee on National Statistics, National Academies of Science, Engineering, and Medicine. (2019). *Workshop on 2020 Census data products: Data needs and privacy considerations*. https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518?

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,* 202–210. https://doi.org/10.1145/773153.773173

Garfinkel, S. L., Abowd, J. M., & Powazek, S. (2018). Issues encountered deploying differential privacy. *Proceedings of the 2018 Workshop on Privacy in the Electronic Society,* 133–137. https://doi.org/10.1145/326

Hotchkiss, M., & Phelan, J. (201
Bureau data in federal funds distri
design for the 21st century. United
Bureau. https://www2.census.g

surveys/decennial/2020/program-management/working-papers/Uses-of-Census-Bureau-Data-in-Federal-Funds-Distribution.pdf

Oberski, D. L., & Kreuter, F. (2020). Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review*, 2(1). https://doi.org/10.1162/99608f92.63a22079

Potok, N. (2019). Deep policy learning: Opportunities and challenges from the Evidence Act. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.77e63f8f

Sullivan, T. A. (2020). Coming To Our Census: How Social Statistics Underpin Our Democracy (And Republic). *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.c871f9e0

U.S. Census Bureau (2019). *2020 Census Memorandum 2019.12: Disclosure Avoidance System design requirement—State resident population and the Federally Affiliated Count Overseas (FACO).* https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_12.html

U.S. Census Bureau. (2020a). *Da*
2020 disclosure avoidance.
https://www2.census.gov/progr
surveys/decennial/2020/progra
management/data-product-planning/disclosure

management/data-product-planning/disclosure-avoidance-system/2020-03-25-data-metrics-2020-da.pdf?#

U.S. Census Bureau. (2020b). *2020 Census operational adjustments due to COVID-19*. https://2020census.gov/en/news-events/operational-adjustments-covid-19.html

U.S. Department of Commerce. (2020). *Statement from Secretary of Commerce Wilbur Ross on the 2020 Census and COVID-19*. https://www.commerce.gov/news/press-releases/2020/03/statement-secretary-commerce-wilbur-ross-2020-census-and-covid-19

READ NEXT • Toward Foundation and Analytics: A Kn Framework for Prof

FOOTNOTES • 5

COMMENTS • 0

Login to discuss

No comments here

EXHIBIT 7

Michael Hawes, U.S. Census Bureau, Differential Privacy and the 2020 Decennial Census at 24 (Mar. 5, 2020)

https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf

Differential Privacy and the 2020 Decennial Census

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

NCSL Webinar
March 5, 2020

Acknowledgements

This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, including: John Abowd, Tammy Adams, Robert Ashmead, Craig Corl, Ryan Cummings, Jason Devine, John Fattaleh, Simson Garfinkel, Nathan Goldschlag, Michael Hawes, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Kyle Irimata, Dan Kifer, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Claudia Molinar, Brett Moran, Ned Porter, Sarah Powazek, Vikram Rao, Chris Rivers, Anne Ross, Ian Schmutte, William Sexton, Rob Sienkiewicz, Matthew Spence, Tori Velkoff, Lars Vilhuber, Bei Wang, Tommy Wright, Bill Yates, and Pavel Zhurlev.

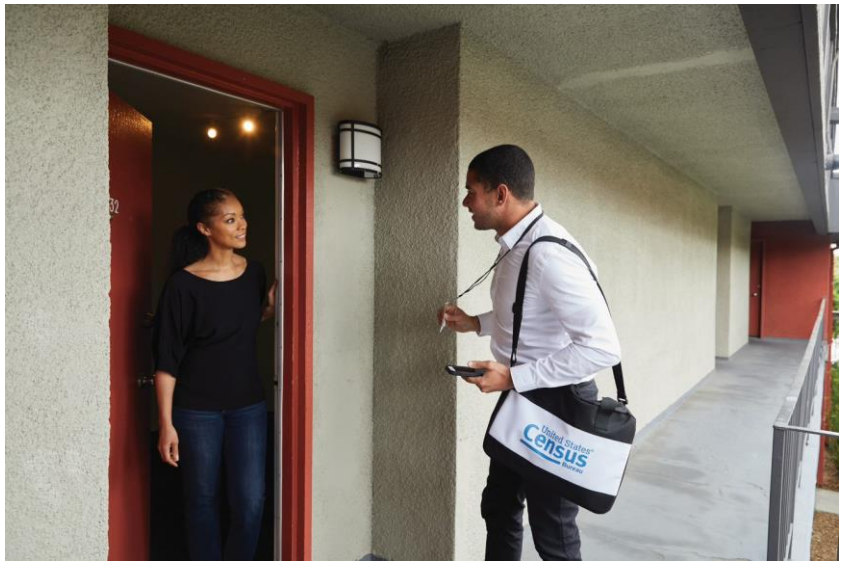
For more information and technical details relating to the issues discussed in these slides, please contact the author at michael.b.hawes@census.gov.

Any opinions and viewpoints expressed in this presentation are the author's own, and do not necessarily represent the opinions or viewpoints of the U.S. Census Bureau.

Our Commitment to Data Stewardship

Data stewardship is central to the Census Bureau's mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



The Privacy Challenge

Every time you release any statistic calculated from a confidential data source you "leak" a small amount of private information.

If you release too many statistics, too accurately, you will eventually reveal the entire underlying confidential data source.

Dinur, Irit and Kobbi Nissim (2003) "Revealing Information while Preserving Privacy" PODS, June 9-12, 2003, San Diego, CA

The Growing Privacy Threat

More Data and Faster Computers!

In today's digital age, there has been a proliferation of databases that could potentially be used to attempt to undermine the privacy protections of our statistical data products.

Similarly, today's computers are able to perform complex, large-scale calculations with increasing ease.

These parallel trends represent new threats to our ability to safeguard respondents' data.

The Census Bureau's Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.

Reconstruction

The recreation of individual-level data from tabular or aggregate data.

If you release enough tables or statistics, eventually there will be a unique solution for what the underlying individual-level data were.

Computer algorithms can do this very easily.

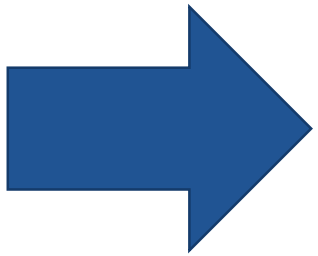
	4						2	
			7					4
1		7	8				5	
			9			3		8
5								
			6		8			
3						4		5
	8	5				1		9
		9		7	1			

Reconstruction: An Example

	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

Reconstruction: An Example

	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7



Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married
36	Female	Black	Married
8	Female	Black	Single
18	Male	White	Single
24	Female	White	Single

This table can be expressed by 164 equations. Solving those equations takes 0.2 seconds on a 2013 MacBook Pro.

Re-identification

Linking public data to external data sources to re-identify specific individuals within the data.

Name	Age	Sex		Age	Sex	Race	Relationship
Jane Smith	66	Female	+	66	Female	Black	Married
Joe Public	84	Male		84	Male	Black	Married
John Citizen	30	Male		30	Male	White	Married

External Data

Confidential Data

In the News

Reconstruction and Re-identification are not just theoretical possibilities…they are happening!

- **Massachusetts Governor's Medical Records (Sweeney, 1997)**
- **AOL Search Queries (Barbaro and Zeller, 2006)**
- **Netflix Prize (Narayanan and Shmatikov, 2008)**
- **Washington State Medical Records (Sweeney, 2015)**
- **and many more…**

Reconstructing the 2010 Census

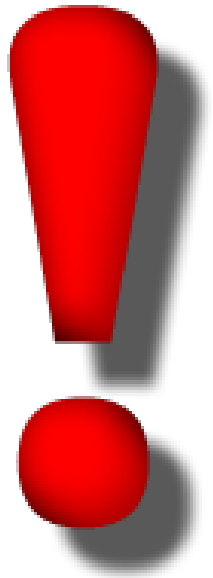
- The 2010 Census collected information on the age, sex, race, ethnicity, and relationship (to householder) status for ~309 Million individuals. (1.9 Billion confidential data points)
- The 2010 Census data products released over 150 billion statistics
- We conducted an internal experiment to see if we could reconstruct and re-identify the 2010 Census records.

Reconstructing the 2010 Census: What Did We Find?

1. On the 309 million reconstructed records, census block and voting age (18+) were correctly reconstructed for all records and for all 6,207,027 inhabited blocks.
2. Block, sex, age (in years), race (OMB 63 categories), and ethnicity were reconstructed:
    1. Exactly for 46% of the population (142 million individuals)
    2. Within +/- one year for 71% of the population (219 million individuals)
3. Block, sex, and age were then linked to commercial data, which provided putative re-identification of 45% of the population (138 million individuals).
4. Name, block, sex, age, race, ethnicity were then compared to the confidential data, which yielded confirmed re-identifications for 38% of the putative re-identifications (52 million individuals).
5. For the confirmed re-identifications, race and ethnicity are learned correctly, though the attacker may still have uncertainty.

The Census Bureau's Decision

- Advances in computing power and the availability of external data sources make database reconstruction and re-identification increasingly likely.
- The Census Bureau recognized that its traditional disclosure avoidance methods are increasingly insufficient to counter these risks.
- To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections.



Differential Privacy

aka "Formal Privacy"

-quantifies the precise amount of privacy risk…

-for all calculations/tables/data products produced…

-no matter what external data is available…

-now, or at any point in the future!

Precise amounts of noise

Differential privacy allows us to inject a precisely calibrated amount of noise into the data to control the privacy risk of any calculation or statistic.

Privacy vs. Accuracy

The only way to absolutely eliminate all risk of re-identification would be to never release any usable data.

Differential privacy allows you to quantify a precise level of "acceptable risk," and to precisely calibrate where on the privacy/accuracy spectrum the resulting data will be.

Data Quality | Bnae Kegouqe
Dada Qualitg | Vrkk Jzcfkdy
Data Qaality | Dncb PrhvBln
Dzte Qvality | Dncb Prtnavy
Dfha Quapyti | Tgta Ppijacy
Tgta Qucjity | Dfha Pnjvico
Dncb Qhulitn | Dzhe Njivaci
Ntue Quevdto | Dzte Privecy
Vrkk Zuhnvry | Dada Privacg
Bnaq Denorbe | Data Privacy

Establishing a Privacy-loss Budget

This measure is called the "Privacy-loss Budget" (PLB) or "Epsilon."

- **$\varepsilon=0$ (perfect privacy) would result in completely useless data**
- **$\varepsilon=\infty$ (perfect accuracy) would result in releasing the data in fully identifiable form**

Epsilon

Comparing Methods

Data Accuracy

Differential Privacy is not inherently better or worse than traditional disclosure avoidance methods.

Both can have varying degrees of impact on data quality depending on the parameters selected and the methods' implementation.

Privacy

Differential Privacy is substantially better than traditional methods for protecting privacy, insofar as it actually allows for measurement of the privacy risk.

Implications for the 2020 Decennial Census

The switch to Differential Privacy will not change the constitutional mandate to apportion the House of Representatives according to the actual enumeration.

As in 2000 and 2010, the Census Bureau will apply privacy protections to the PL94-171 redistricting data.

The switch to Differential Privacy requires us to re-evaluate the quantity of statistics and tabulations that we will release, because each additional statistic uses up a fraction of the privacy-loss budget (epsilon).

Demonstrating Privacy, Assessing and Improving Accuracy

The DAS Team's priorities over Fall 2019 were:

- **To scale up the DAS to run on a (nearly) fully-specified national histogram**
- **To demonstrate that the DAS can effectively protect privacy at scale**
- **To permit the evaluation and optimization of the DAS for accuracy and "fitness for use"**

These initiatives were largely successful, but much more work needs to be done over the remainder of this year.

The engagement and efforts of our data users have been enormously helpful in helping to identify and prioritize this remaining work.

Harvard Data Science Review Symposium

Held at Harvard University on October 25, 2019

Evaluated the DAS using public 1940 Census data

Assessments by teams of data users from:

- **NORC at the University of Chicago** – Sampling Efficiency and Funding Allocations
- **IPUMS at the University of Minnesota** – Racial Residential Segregation
- **W.E. Upjohn Institute for Employment Research** – Scrubbed Segregation

Committee on National Statistics Workshop

December 11-12, 2019

Evaluation of 2010 Census data run through a preliminary version of the 2020 DAS

Data user assessments and findings on DAS implications for:

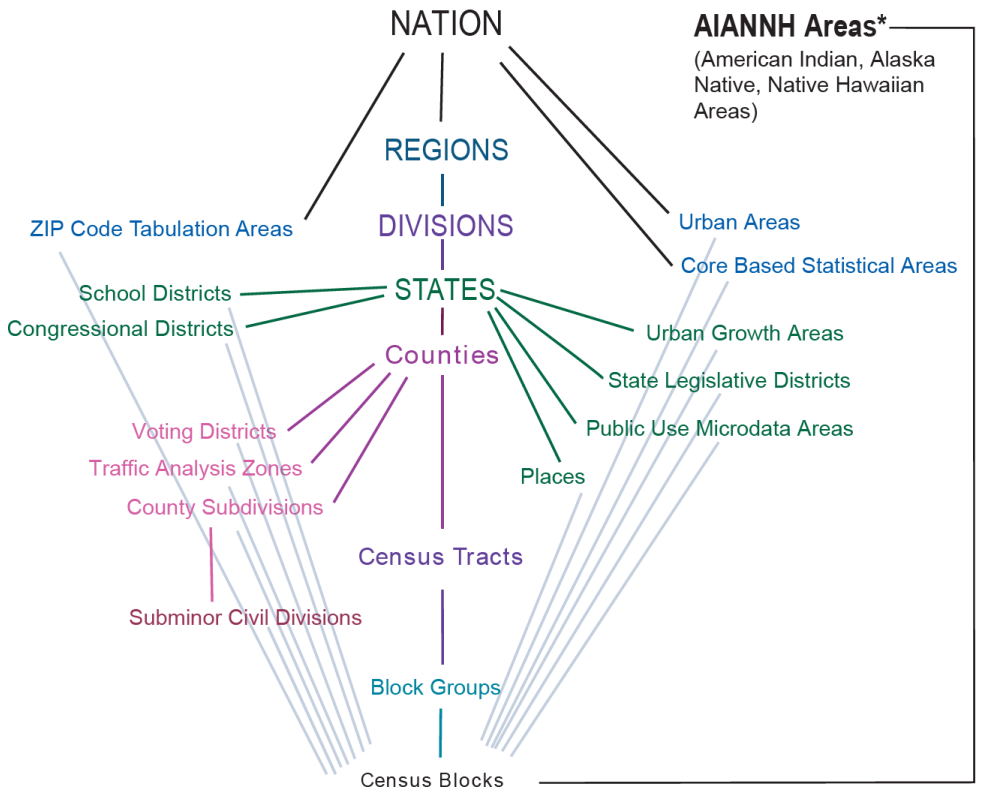
- **Redistricting and related legal use cases**
- **Identification of rural and special populations**
- **Geospatial analysis of social/demographic conditions**
- **Delivery of government services**
- **Business and private sector applications**
- **Denominators for rates and baselines for assessments**

What We've Learned: Accuracy

- **The October vintage of the DAS falls short on ensuring "fitness for use" for several priority use cases.**
- **There are two sources of error in the TopDown Algorithm (TDA):**
  - Measurement error due to differential privacy noise
  - Post-processing error due to statistical inference creating non-negative integer counts from the noisy measurements
- **Post-processing error tends to be much larger than differential privacy error**
- **Positive bias in small counts/negative bias in large counts is the result of**
  - Invariants
  - Post-processing error specifically introduced by our Non-negative Least Squares (L2) optimization routine
- **Improving post-processing is not constrained by differential privacy**
- **Current initiatives include incorporating legal and political geographies into the geographic spine and adopting a multi-phase approach to post-processing**

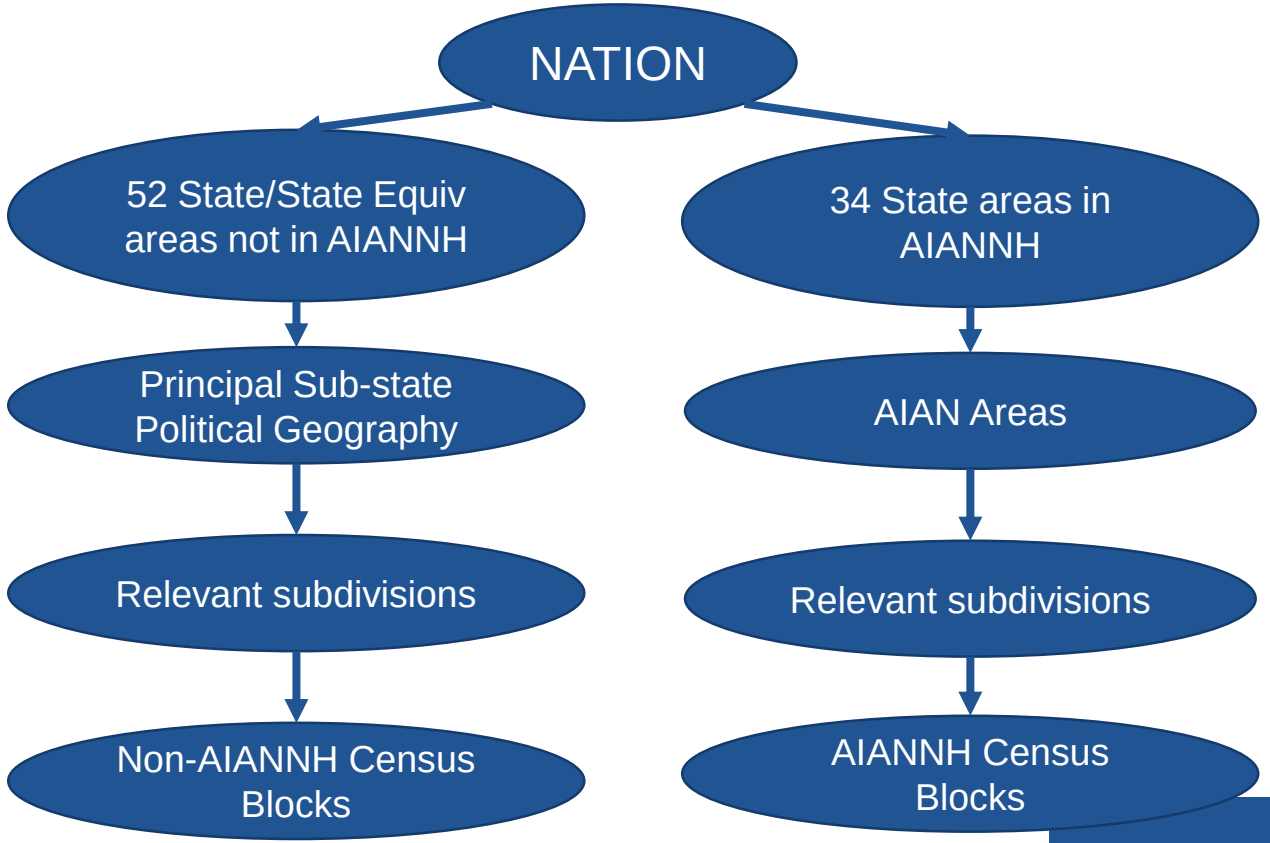
Revising Geographical Hierarchy to address count accuracy for AIANNH and INCPLACE/CDPs

Old Hierarchy:



2020CENSUS.GOV

New Approach (work in progress):



Shape your future START HERE >

United States Census 2020

Making population counts more accurate.

Nearly all of the error in the 2010 Demonstration Data Products came from post-processing, not from differential privacy.

Old approach:

Single-pass post-processing:

- **Optimize accuracy for ~1.2M histogram cells (2010 DDP used only ~400,000 cells).**
- **All cells must be integers**
- **All cells must be ≥0**
- **All margins must satisfy adding up constraints within and between levels of the geographic spine**
- **All invariants and structural zeros must hold exactly**

New Approach (work in progress):

Multi-pass post-processing:

- **First pass: compute total population and GQ populations**
- **Second pass for redistricting file (total pops constrained to first pass values)**
- **Third pass for population-estimates program. 3M tabs. (counts constrained to second pass values)**
- **Fourth pass: rest of DHC-H and DHC-P (counts constrained to values from passes above)**

Making population counts more accurate.

A set of metrics are being developed based on use cases and stakeholder feedback. The metrics will allow the public to see the improvements that are made leading up to the finalization of the TopDown Algorithm (TDA).

The selected metrics will:

- **Be straightforward and easy to interpret**
- **Reflect input from external data users;**
- **Show differences between major DAS runs and publicly available 2010 tabulations**
- **Provide accuracy, bias, and outlier information for basic demographic tabulations**
- **Provide accuracy, bias, and outlier information for categories of use cases**

These metrics will inform data users of accuracy improvements we are able to make while also informing their ongoing engagement throughout the remaining work.

Questions?

Disclosure Avoidance and the 2020 Census Website
https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

301-763-1960 (Office)
michael.b.hawes@census.gov

EXHIBIT 8

John M. Abowd, Modernizing Disclosure Avoidance: A Multipass Solution to Post-processing. Error, The Census Bureau, (June 18, 2020)

https://www.census.gov/newsroom/blogs/research-matters/2020/06/modernizing_disclosu.html

Modernizing Disclosure Avoidance: A Multipass Solution to Post-processing Error

June 18, 2020

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY AND DR. VICTORIA A. VELKOFF, ASSOCIATE DIRECTOR FOR DEMOGRAPHIC PROGRAMS

In our last blog [https://www.census.gov/newsroom/blogs/research-matters/2020/03/modernizing_disclosu.html] , we discussed the feedback we received from the data user community about demonstration data released last fall that were produced using the interim version of the 2020 Disclosure Avoidance System (DAS). It was clear that the fall 2019 version of the DAS TopDown Algorithm (TDA) introduced unacceptable amounts of error and distortion into statistics used for many important use cases. In that blog, we also discussed our ongoing plans to improve the algorithm to address and mitigate this error.

The team responsible for developing the DAS uses an *agile* development approach, which implements improvements to the system in a series of four-week development *sprints*. During the sprint that concluded in March 2020, we began implementing changes to address those issues. The most notable change involved how the TDA converts the formally private noisy tabulations taken from the confidential data into the non-negative integer counts that will be published, an operation that we call “post-processing.”

Previously, the TDA conducted the post-processing of all of the statistics for a particular geographic level at the same time. Unfortunately, as we saw in the demonstration data the TDA had difficulty accurately performing this optimization when there were large quantities of statistics with zeros or very small values processed at the same time. The result was distortions in the data that effectively moved individuals from high- to low-density populations (e.g., from cities to rural areas, or from larger race groups to smaller race groups).

During the March sprint, we implemented a change to the algorithm design to address and mitigate this issue. Now, the TDA conducts the post-processing in a series of passes through all the geographic levels.

At the national level, the state level, and finally at each lower level of geography, the first pass of the algorithm solely determines the population count for each unit within that geographic level (e.g., for all census tracts within a county).

Once those total population counts are determined, the second pass of the algorithm processes just the statistics necessary to produce the redistricting data (also known as the Public Law 94-171 data file), constraining those statistics to the sum of the population counts determined in the first pass.

The third pass through the algorithm then processes the core statistics necessary to support population by age, sex, and broad race/ethnicity categories for the demographic analyses that underlie the Population Estimates Program. Third-pass statistics are constrained to the sum of the statistics produced for the redistricting data.

A final pass through TDA processes the remainder of the statistics necessary for the Demographic and Housing Characteristics files and the Demographic Profiles, constraining these values to the sum of the ones produced in the third pass.

At the same time, the team examined options for improving the accuracy of population counts for legal and political entities, including American Indian, Alaska Native and Native Hawaiian areas, minor civil divisions, incorporated places, etc. Census Bureau geography experts determined the optimal geographic entities to prioritize for accuracy within each state based on knowledge gathered from decades of preparing geographic hierarchies in support of state and local government objectives.

While the DAS geographic hierarchy itself was not modified, the way the total population query was handled in the latest version of the DAS demonstrates that population accuracy is now controlled by the privacy-loss budget directly and not by errors induced by post-processing.

Identifying and prioritizing future improvements to the DAS requires ongoing dialogue with our data users. To facilitate that dialogue, we are committed to demonstrating how much each major change to the TDA design improves accuracy and “fitness for use” of the resulting data for many of the priority use cases identified by our data users. As we previously discussed, the Census Bureau has developed a comprehensive suite of error measures [https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html] to use to evaluate the improvements we are making to the algorithm throughout 2020. We are consulting with a group of experts [https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518] identified by the Committee on National Statistics to ensure that these are the appropriate accuracy measures to use. We also welcome input from our other data users. You can send suggestions and feedback to <2020DAS@census.gov [mailto:2020DAS@census.gov] >.

On May 27, we published Detailed Summary Metrics [https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-05-27-data-metrics-tables.xlsx] , which are an evaluation of a full run of TDA from the March sprint that incorporated our new multipass approach to post-processing. Comparing the accuracy of this data set to baseline measures [https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-25-data-metrics-tables.xlsx] run on the 2010 Demonstration Data Products shows we have substantially reduced the error associated with population counts in the demonstration data.

For example, in the 2010 Demonstration Data Products [https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html] , the total population count for the average county was off by approximately 82 people (0.78%).With the algorithmic improvements we implemented in March, that error dropped to just 16 people (0.14%). These improvements are also observable at lower levels of geography. In the demonstration product run, total population for the average census tract was off by almost 26 people; now that error has been reduced to just 14.5 people. At the block level, error in the population for the average urban census block dipped from 9.2 to 7.7 people.

These accuracy improvements come without any reduction in the strength of the privacy guarantee. That is, the privacy-loss budget for both DAS runs held constant, so the observed improvements are directly attributable to improvements in our post-processing algorithm. More work remains to be done, however, and we look forward to sharing our progress with you through this blog and additional releases of the accuracy measures on future runs of the DAS.

This entry was posted on June 18, 2020 and filed under Disclosure Avoidance [/newsroom/blogs/research-matters.html/category/Topic/research/disclosure-avoidance] and Population [/newsroom/blogs/research-matters.html/category/Topic/Their Population].

EXHIBIT 9

Andy Beveridge, Sixteen States Sue to Block Census Bureau Data Privacy Method (Apr. 19, 2021)

https://www.socialexplorer.com/blog/post/sixteen-states-sue-to-block-census-bureau-data-privacy-method-11411

Sixteen States Sue to Block Census Bureau Data Privacy Method

MONDAY, APR 19, 2021

The Census Bureau plans to release the state population counts used to apportion Congress by April 30. In August, it will release the first wave of redistricting data, which includes population counts and racial/ethnic distributions for all Census geographies — states, cities, counties, towns, villages, voting districts, tracts, block groups, and blocks. Data will be distorted to ensure privacy, but the process is likely to make the data much less useful for many purposes, including redistricting.

Test versions of the new methods applied to the 2010 data showed that serious inaccuracies were introduced, making the data much less useful and suspect for even the simple task of reporting the population and racial and ethnic distributions for small geographies. In early March, Alabama sued to block the distortion. It was joined last week by 16 states. A court hearing is scheduled in two weeks, and the outcome may be immediately appealed to the U.S. Supreme Court.

Since the plans to distort the data were announced in spring of 2019 (and especially since a test version based upon the 2010 Census was released in October 2019), many users have raised serious questions about the data and its implementation, made without consulting states.

According to the Census Bureau, a new "accuracy metric" is being implemented for the next test version to be released around April 30. The Census Bureau says the metric will "ensure that the largest racial or ethnic group in any geography entity with a total population of at least 500 people is accurate to within 5 percentage points of their enumerated value at least 95% of the time."

This approach immediately raises several questions:

1. Within the entities defined (presumably by assemblies of Census blocks), how will the second-largest racial or ethnic groups be affected?. If the groups are roughly the same size, it may be important to define the majority group for voting rights enforcement. And how will this affect outliers that are not within 5 percentage points?
2. What entities used for voting (e.g., precinct-like entities, as well as other Census geographies) would likely be ignored because they have fewer than 500 people? We have presented a set of tables that shows the number and percentages of counties, places, minor civil divisions, tracts, block groups, voting tabulation districts, and blocks with fewer than 500 people. (We used 2010 Census data for voting tabulation districts and blocks; for all other geographies, we used the 2015-19 American Community Survey).

These tables show most blocks have fewer than 500 people. Roughly one-fifth of voting tabulation districts have fewer than 500 people. A substantial number of other geographies – other than counties, tracts, and block groups – also have fewer than 500 people.

All Places Less than 500

The FREQ Procedure

Pop_500_or_more	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	9537	32.82	9537	32.82
YES	19524	67.18	29061	100

Incorporated Places Less Than 500

The FREQ Procedure

Pop_500_or_more	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	6025	30.9	6025	30.9
YES	13474	69.1	19499	100

All MCDS Less Than 500

The FREQ Procedure

Pop_500_or_more	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	10074	28.61	10074	28.61
YES	25136	71.39	35210	100

All Counties Less Than 500

The FREQ Procedure

Pop_500_or_more	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	8	0.25	8	0.25
YES	3134	99.75	3142	100

All Tracts Less Than 500

The FREQ Procedure

Pop_500_or_more	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	245	0.34	245	0.34
YES	72165	99.66	72410	100

All Block Groups Less Than 500

The FREQ Procedure

Pop_500_or_more	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	8084	3.73	8084	3.73
YES	208600	96.27	216684	100

All VTDs Less Than 500

The FREQ Procedure

Pop_500_or_more	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	30865	17.83	30865	17.83
YES	142213	82.17	173078	100

All 2010 Census Blocks Less Than 500

The FREQ Procedure

Pop_500_or_more	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	6206948	99.19	6206948	99.19
YES	50999	0.81	6257947	10

Once the new demonstration product and metrics are released, we will continue to analyze their impact using tools provided by the Census Bureau. We will also continue to monitor the outcome of the court case attempting to block the use of the new distorting privacy method.

Author: Andy Beveridge

EXHIBIT 10

Aref N. Dajani et al., Presentation to Census Scientific Advisory Committee, The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau (Sept. 2017)

https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf

The modernization of statistical disclosure limitation at the U.S. Census Bureau

Aref N. Dajani[1], Amy D. Lauger[1], Phyllis E. Singer[1], Daniel Kifer[2], Jerome P. Reiter[3], Ashwin Machanavajjhala[4], Simson L. Garfinkel[1], Scot A. Dahl[6], Matthew Graham[7], Vishesh Karwa[8], Hang Kim[9], Philip Leclerc[1], Ian M. Schmutte[10], William N. Sexton[11], Lars Vilhuber[7, 11], and John M. Abowd[5]

[1] Center for Disclosure Avoidance Research, U.S. Census Bureau, *firstname.m.lastname@census.gov*
[2] Department of Computer Science and Engineering, Penn State University, dkifer@cse.psu.edu
[3] Department of Statistical Science, Duke University, jerry@stat.duke.edu
[4] Department of Computer Science, Duke University, ashwin@cs.duke.edu
[5] Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau, John.Maron.Abowd@census.gov
[6] Economic Statistical Methods Division, U.S. Census Bureau, Scot.Alan.Dahl@census.gov
[7] Center for Economic Studies, U.S. Census Bureau, *firstname.m.lastname@census.gov*
[8] Department of Statistics, Harvard University, vkarwa@seas.harvard.edu
[9] Department of Mathematical Sciences, University of Cincinnati, hang.kim@uc.edu
[10] Department of Economics, University of Georgia, schmutte@uga.edu
[11] Labor Dynamics Institute, Cornell University, {wns32,lv39}@cornell.edu
[12] Economic Statistical Methods Division, U.S. Census Bureau, Katherine.J.Thompson@census.gov

Abstract: Most U.S. Census Bureau data products use traditional statistical disclosure limitation (SDL) methods such as cell or item suppression, data swapping, input noise infusion, and censoring to protect respondents' confidentiality. In response to developments in mathematics and computer science since 2003, the Census Bureau is developing formally private SDL methods to protect its data products. These methods will provide mathematically provable protection to respondents and allow policy makers to manage the tradeoff between data accuracy and privacy protection—something previously done by technical staff. The Census Bureau's OnTheMap tool is a web-based mapping and reporting application that shows where workers are employed and where they live. OnTheMap was the first production deployment of formally private SDL in the world. Recent research for OnTheMap has incorporated formal privacy guarantees for businesses to complement the existing formal protections for individuals. Research is underway to improve the disclosure limitation methods for the 2020 Census of Population and Housing, the American Community Survey, and the 2017 Economic Census. For each of these programs, we are developing models to create synthetic microdata, from which we can create aggregated estimates. There are many challenges in adopting formally private algorithms to datasets with high dimensionality and the attendant sparsity. We are also developing approaches for gauging the synthetic data's accuracy and usefulness for specific applications. In addition to formally private methods that allow senior executives to set the privacy-loss budget, our implementation will feature adjustable "sliders" for allocating the privacy-loss budget among related sets of tabular summaries. The U.S. Census Bureau will implement the settings for the privacy-loss budget and these sliders using recommendations from the Data Stewardship Executive Policy Committee, as was done in the 2000 and 2010 Censuses.

1 Overview: Disclosure Limitation at the U.S. Census Bureau Today

The U.S. Census Bureau views disclosure limitation not just as a research interest, but as an operational imperative. The Bureau's hundreds of surveys and censuses of households, people, businesses, and establishments yield high quality data and derived statistics only if the Bureau maintains effective data stewardship and public trust.

The Bureau has traditionally used statistical disclosure limitation (SDL) techniques such as top- and bottom-coding, suppression, rounding, binning, noise-infusion, and sampling to reserve the confidentiality of respondent data. The Bureau is currently transitioning from these SDL methods to modern SDL techniques based on formally private data publication mechanisms.

1.1 Legal Requirements

The Census Bureau collects confidential information from U.S. persons and businesses under the authority of Title 13 of the U.S. Code. Once collected, the confidentiality of that data is protected specifically by 13 USC 9, which prohibits:

(i) Using the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or
(ii) Making any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or
(iii) Permitting anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual records.

Some publications are further protected by Title 26 of the U.S. Code, which also protects the federal tax information (FTI) used by the Bureau in the preparation of statistical products, primarily from businesses. Additionally, the Department of Commerce (2017), in which the Bureau is housed, has issued directives regarding the protection of personally identifiable information (PII) and business identifiable information (BII). These directives largely mirror those issued by other government agencies and prohibit release of information that can be used "to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."

1.2 Current methods supporting statistical disclosure limitation (SDL)

Currently, the Bureau primarily uses information reduction and data perturbation methods to support SDL (Lauger et al., 2014). Information reduction methods include swapping, top- and bottom-coding, suppression, rounding or binning, and sampling collected units for release in public use microdata files. Current data perturbation methods include swapping, noise infusion, and partially and fully synthetic database construction. The current approach starts with the premise that there are specific data elements that must be protected (e.g., a person's income). A technical analyst choses an

approach from the assortment of available SDL methods that is likely to protect the data without resulting in too much damage to the published data accuracy.

These ad hoc approaches do not offer formal guarantees of data confidentiality. That is, a person's income may be suppressed in a cell, but it may be possible to reconstruct that person's income by combining information published elsewhere within the statistical tables; that is, without using any external data.

1.3 Formal privacy approaches

Formal privacy methods take a different approach to protecting confidential information. Instead of starting with a list of confidential values to protect and an ad hoc collection of protection mechanisms, the formal approach starts with a mathematical definition of privacy. Next, it implements mechanisms for publishing *queries* based on the confidential data that are provably consistent the formal privacy definition. Thus, the tables released by statistical agency are actually modeled as a series of queries applied to the confidential data. Surrogates for public use microdata samples (PUMS) files can also be generated in this manner: instead of sampling the actual respondent data, the queries are used to create formally private synthetic data. This is done by first modeling the confidential data, then using the model to generate synthetic data, as discussed below.

Differential privacy (Dwork et al., 2006) is the most developed formal privacy method. It begins by specifying the structure of the confidential database to be protected, D. In computer science, this is called the database schema, in statistics the sample space. Two databases, D_1 and D_2, with the same schema are neighbors if the appropriately defined distance between them is unity. Leaving the technical details aside, say $|D_1 - D_2| = 1$. The universe of tables to be published from D is modeled as a set of queries on D, say Q. An element of Q, say q, is a single query on D. A randomized algorithm, A, takes as inputs D, q, and an independent random variable. The output of $A(D, q)$ is the statistic to be published, say S, which a measureable set in the probability space defined by the independent random variable, say B. A randomized algorithm A for a publication system for releasing all of the queries in Q is ε-differentially private if, for all D_1 and D_2, with the same database schema and $|D_1 - D_2| = 1$, for all $q \in Q$, and for all $S \in B$

$$\Pr[A(D_1, q) \in S] \leq e^{\varepsilon} Pr[A(D_2, q) \in S].$$

The probability is defined by the independent random variable that is used by the algorithm A, and not by the probability of observing any database D with the allowable schema (likelihood function in statistics).

There are alternative ways to define adjacent databases. For example, one method considers the databases adjacent if the record of a single person is added or removed from the database. Alternatively, the value of a single data item on a single record can be changed. Differential privacy is the mathematical formalization of the intuition that a person's privacy is protected if the statistical agency produces its outputs in a manner insensitive to the presence or absence of that persons data in the confidential database.

In differential privacy, the value ε is the measure of privacy loss or confidentiality protection. If $\varepsilon = 0$, then the two probability distributions in the definition always produce exactly the same answer from neighboring inputs—there is no difference in the output of algorithm A when given adjacent database inputs. Since the definition applies to the universe of potential inputs, and all neighbors of those inputs, all databases therefore produce exactly the same answer. Thus, the value $\varepsilon = 0$ guarantees no privacy loss at all (perfect confidentiality protection), but no data accuracy, since it is equivalent to encrypting the statistic S. In contrast, when $\varepsilon = \infty$, there is no confidentiality protection at all—full loss of privacy, but the statistics S are perfectly accurate (identical to what would be produced directly from the confidential input database). Thus, ε can be thought of as the *privacy-loss budget* for the publication of the queries in Q: the amount of privacy that individuals must give up in exchange for the accuracy that can be allowed in the statistical release.

Varying the privacy-loss budget allows us to move along a privacy-accuracy Production Possibilities Frontier (PPF) curve, as it is known in the economics literature, or along the Receiver Operating Characteristics (ROC) curve, as it is known in the statistics literature (Abowd and Schmutte 2017). The curve constrains the aggregate disclosure risk that confidential data might be jeopardized through any feasible reconstruction attack, given all published statistics for any attacker. This budget is the worst-case limit to the inferential disclosure of any identity or item. In differential privacy, that worst case is over all possible databases with the same schema for all individuals and items.

The privacy-loss budget applies to the combination of *all* released statistics that are based on the confidential database. As a result, the formal privacy technique provides protection into the indefinite future and is not conditioned upon additional data that the attacker may have.

To prove that a privacy-loss budget is respected, one must quantify the privacy-loss expenditure of each publication or published query. The collection of the algorithms considered altogether must satisfy the privacy-loss budget. This means that the collection of algorithms used must have known composition properties.

Because the information environment is changing much faster than when traditional SDL techniques were developed, it may no longer be reasonable to assert that a product is empirically safe given best-practice disclosure limitation prior to its release. Formal privacy models replace empirical disclosure risk assessment with designed protection. Resistance to all future attacks is a property of the design.

Differential privacy, the leading formal privacy method, is robust to background knowledge of the data, allows for sequential and parallel composability and allows for arbitrary post-processing edits. Differential privacy's proven guarantees hold even if external data sources are published or released later. Other formal privacy methods quantify the privacy loss that can also be mathematically established and proven, but with more constrained properties (Haney et al., 2017).

2 Expanding privacy protection for OnTheMap

Randomized response, a survey technique invented in the 1960s, was the first differentially private mechanism implemented by any statistical agency, although it was not a conscious decision, and the technique is difficult to adapt to modern survey collection methods (Wang et al., 2016).

The first production application of a formally private disclosure limitation system by any organization was the Census Bureau's OnTheMap (residential side only), a geographic query response system for studying residence and workplace patterns.

The Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES), the data used by OnTheMap, is a partially synthetic dataset that describes geographic patterns of jobs by their employment locations and residential locations as well as the connections between the two locations (U.S. Census Bureau, 2016). A job is counted if a worker is employed with positive earnings during the reference quarter and in the quarter prior to the reference quarter. These data and marginal summaries are tabulated by several categorical variables. The origin-destination (OD) matrix is made available by ten different "labor market segments". The area characteristics (AC) data–summary margins by residence block and workplace block–contain additional variables including age, earnings, and industry. The blocks are defined in terms of 2010 Census blocks, defined for the 2010 Census of Population and Housing. The input database is a linked employer-employee database, and statistics on the workplaces (Quarterly Workforce Indicators: QWI) are protected using noise infusion together with primary suppression (Abowd et al., 2009, 2012).

For OnTheMap and the underlying LODES data, the protection of the residential addresses is independent of the protection of workplaces. Protection of worker information is achieved using a formal privacy model (Machanavajjhala et al., 2008); work is in progress to protect workplaces using formal privacy as well (Haney et al., 2017).

3 SDL methods supporting the 2020 Census of Population and Housing

The Census Bureau is testing the feasibility of producing differentially private tabulations of the redistricting data (PL94-171) for the 2018 End-to-End Test. It is currently in the process of algorithm development and obtaining the cloud computing environment necessary to scale the research to the requirements of the Census of Population and Housing. For the full 2020 Census, the Bureau will extend the methods used for the 2018 End-to-End test to the tabulations in Summary File 1.

The differentially private tabulations for the 2020 Census will support the following products:

- **Public Law (PL) 94-171** for redistricting,
- **Census Summary File (SF) 1** for demographic and housing counts, and

- **Geographical Hierarchy** from the national to the block level, exploiting parallel composition to efficiently use the privacy-loss budget.

By 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,
- Number of vacant housing units, and
- Number of householders, which is equal to the number of occupied housing units.

Key disclosure limitation challenges include:

1. Ensuring consistency by respecting the unaltered counts enumerated above,
2. Respecting joins; e.g., to group people into households,
3. Large memory/time requirements for explicitly stored universes and well-understood low-dimensional approximations,
4. Difficulty detecting coding errors, particularly as pertains to verifying privacy-loss guarantees,
5. Communicating analytical results clearly to, and in a format useful for, policy makers,
6. A lack of high-quality usage data from which to infer relative importance of data products, and
7. Determining how much of the privacy-loss budget should be spent per household; e.g., whether it should be proportional to household size.

The 2010 and 2000 Censuses of Population and Housing applied SDL in the form of record swapping, but this fact was not always obvious to data users. The actual swapping rate is confidential, as is the impact that swapping had on overall accuracy. Throughout each decade, the Census Bureau also conducts special tabulations of small geographic areas such as towns. Those tabulations also impact privacy, and they also undergo SDL.

Key policy-related challenges include:

1. Communicating the global disclosure risk-data accuracy tradeoff effectively to the Disclosure Review Board and Data Stewardship Executive Policy Committee so that they can set the privacy-loss budget and the relative accuracy of different publications,
2. Providing effective summaries of the social benefits of privacy vs. data accuracy, so that DESP, in particular, can understand how the public views these choices.

4 SDL methods supporting the American Community Survey (ACS)

The American Community Survey (ACS) is the successor to the long form survey of the Census of Population and Housing. The housing unit survey includes housing,

household, and person-level demographic questions about a broad range of topics. There is a separate questionnaire for those residing in group quarters. The Bureau sends this survey to approximately 3.5 million housing units each year and receives approximately 2.5 million responses. Weighted adjustments account for nonresponse, in-person interview subsampling, and raking to pre-specified population controls. The ACS sample is usually selected at the tract level and is designed to allow reliable inferences for small geographic areas and for subpopulations, when averaged across five years. ACS sampling rates vary across tracts. On average, a tract will have approximately thirty-five housing units and ninety people in the returned sample.

The Bureau releases one-year and five-year ACS data products. Five-year tables are released either by block group or by tract. One-year tables have been released only for geographies containing at least 65,000 people. A recent DRB decision allowed some one-year tables to be released for areas of at least 20,000, due to the termination of the three-year data products.

The feasibility of developing formally private protection mechanisms given current methodological and computational constraints, the large number of ACS variables, and the desire for small area estimates is undemonstrated. The Bureau is actively pursuing this research, seeking to leverage advances from other data products. As an intermediate step, the Bureau is experimenting with non-formally private synthetic data using statistical models to replace the current SDL methods.

Key disclosure avoidance challenges include:

1. **High dimensionality:** there are roughly two hundred topical module variables with mixed continuous and categorical values,
2. **Geography**, with estimates needed at the Census tract level,
3. **Preserving associations** among variables across people in the same household,
4. **Outliers** in the economic variables,
5. **Dealing with weighting** due to sampling and nonresponse adjustment.

These challenges stem from high dimensionality combined with small sample sizes. Small geographies and sub-populations are important for data users. Tract-level and even block group-level data are critical for many applications, including the ballot language determinations in Section 203 of the Voting Rights Act. In addition to legislative districts, many special geographies published by the Census Bureau, including cities and school districts, are dependent upon small component geographies.

The large margins of error for small geographies allow some scope for introducing error from SDL without significantly increasing total survey error. Modelling can introduce some bias for massive decreases in variances by borrowing strength from correlations.

The research team is considering the following approach:

1. Build a chain of models, simulating each variable successively given the previous synthesized variables (Raghunathan et al., 2001)

2. Build a formally private version of these models, if feasible.
3. Create microdata samples from these models.
4. Create tables from these microdata samples.

Validation servers, verification servers, and access to the FSRDCs may be the solution for research questions for which the modernized SDL approach leads to reasonable uncertainty regarding the suitability of published data for a particular use. An advantage of the methods being tested for both the 2020 Census and the ACS is that they permit quantification of the error contributed by the SDL; hence, the inferences from the published data are correct. Their suitability for use in a particular application can, therefore, be assessed without reference to the confidential data. This property of the modernized SDL provides a means for applying objective criteria to a researcher's claim that the published data are unsuitable for a particular use.

5 SDL research supporting the 2017 Economic Census

Every five years the Bureau sends survey forms to nearly four million U.S. business establishments, broadly representative of the complete U.S. geography and most private industries, to conduct the Economic Census.

The Bureau defines an *establishment* as a specific economic activity conducted at a specific location. The Bureau asks companies to file separate reports when operating at different locations and when multiple lines of activity are present at a given location. The Economic Census is thus a mixture of a complete enumeration for certain types of businesses, and sampling of other types.

The Economic Census collects information from sampled establishments on the revenue obtained from product sales ("products") in any given industry. Establishments can report values from a wide variety of potential products. The reported product values are expected to sum to the total receipts reported earlier in the questionnaire. Often, product descriptions are quite detailed, and many products are mutually exclusive. Consequently, legitimate missing values occur frequently. Good predictors such as administrative data and other survey data are available for variables such as revenue, payroll, and employment, but auxiliary data are not available for the other items.

The key challenge that the development team will focus on is the disclosure limitation process for North American Product Classification System (NAPCS) product estimates that are new to 2017. The current plan is to release product and product-by-industry tabulations that satisfy predetermined privacy and reliability constraints and to release supplemental synthetic industry-level microdata files, pending the outcome of the research discussed below.

Beginning in 2017, an interdisciplinary team at the Census Bureau partnered with academic colleagues to evaluate the feasibility of developing synthetic industry-level microdata comprising general statistics items and selected products. Specific products

may differ by industry and the level of model estimation (industry, industry by state) will need to be determined in the course of the research.

Kim, Reiter, and Karr (2016) present methods of developing synthetic data on historic Economic Census data from the manufacturing sector. The goal is to extend their multivariate joint model to accommodate additional Economic Census industries, modifying them as the research indicates. There are other publications already approved for the 2017 Economic Census; hence, the synthetic data must satisfy additional constraints—specifically the preservation of published margins. The proposed methods allow for multiple imputation variance estimation. It has not been determined whether the multiple imputation variance estimates for the synthetic data will need to approximately match the published variance estimates.

In addition to developing usable datasets, there is an additional goal of teaching users to use synthetic data to produce their own tabulations and conduct their own analyses. The team thus needs to consider usage and analysis by outside users.

6 Challenges and meetings those challenges

In differential privacy, the commonly used flattened histogram representation of the universe is calculated as the Cartesian product of all potential combinations of responses for all variables. This representation is often orders of magnitude larger than the total population even when structural zeroes (impossible combinations of values of variables, such as grandmothers three years of age) are imposed.

Policy makers, including the Data Stewardship Executive Policy Committee of the Census Bureau, must have enough information about the privacy-loss/data accuracy trade-off to make an informed decision about ε, and its allocation to different tabular summaries. In some cases, the chosen amount of noise infusion from differential privacy may limit the suitability for use of the published statistics to more narrowly defined domains than has historically been true.

The strategy for producing the tabular summaries is to supply the official tabulation software with formally private synthetic data that reproduce all of the protected tabulations specified in the redistricting and summary file requirements. In generating high quality synthetic microdata, one needs to consider integer counts, non-negativity, unprotected counts (e.g., voting age population), and structural zeroes.

To execute this approach, the Bureau needs generic methods that will work on a broader range of datasets. In addition, it may be difficult to find meaningful correlations that are not represented in the model. To address this, the model must anticipate the analysis that many downstream users might conduct. As a result, better model-building tools are needed, as well as generic tools for correlating arbitrary models with the ones used to build the synthetic data.

Reproducible-science methods will be required to use synthetic data effectively.

Data are often collected with a complex sample design with considerable missing data and in panels of longitudinal data. Research is ongoing to ensure that weighted, longitudinal analysis using differentially private data will continue to produce "good results and good science" to the data users.

7 Approaches to gauge data accuracy and usefulness

There are multiple methods to establishing data accuracy, also known as analytical (or inference) validity. Machanavajjhala et al. (2008) conducted experiments comparing differentially private synthetic data to the actual data for OnTheMap. They saw value in coarsening the domain to limit the number of "strange fictitious commuting patterns." Karr et al. (2006) and Drechsler (2011) advocate calculating confidence interval overlaps for parameters of interest, whether univariate, bivariate, or multivariate.

There is value in calculating all such metrics described above for parameter estimates calculated from:

- non-perturbed data (exact counts) where we expect parity.
- parameters estimates that were not captured in the joint distributions modeled in the synthetic data, where one would not expect to uncover comparable results.

Disclosure limitation is a technology. It shows the relationship between privacy loss, which is considered a public "bad", and data accuracy, which is considered a public "good". A differentially private system can publish extremely disclosive data. This happens if the privacy-loss budget is set very high. The extremely disclosive data are also very accurate. That is, inferences based on these data are nearly identical to those based on the confidential data. But extremely disclosive, albeit formally private, data also permit a very accurate reconstruction of the confidential data relative to the reconstruction possible with smaller privacy-loss budgets.

The teams at the Census Bureau working on formal privacy methods for statistical disclosure limitation have been charged by DSEP with developing technologies with adjustable parameters to control the privacy loss and data accuracy during implementation. Those technologies will be summarized with a variety of supporting materials. The Disclosure Review Board will make a recommendation regarding the appropriate formal privacy technology and parameter settings, including the privacy-loss parameter ε. The Data Stewardship Executive Policy Committee will review that recommendation and forward its recommendation to the Director. The published data will implement the recommendations of DSEP, as they have for the past two decennial censuses. Although more explicit than in previous censuses, this is the same chain of recommendation and approval that was used in 2000 and 2010.

This transition to innovation involves significant retooling of methods for the Census Bureau's career mathematical statisticians, IT specialists, project and process managers, and internal stakeholders. This transition will help the Census Bureau lead similar innovation across the U.S. Federal Government and beyond.

8 References

Abowd, John M. and Ian Schmutte (2017). *Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods*. Under review. http://digitalcommons.ilr.cornell.edu/ldi/37/.

Abowd, John M., R. Kaj Gittings, Kevin L. McKinney, Bryce Stephens, Lars Vilhuber, and Simon D. Woodcock (2012). *Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures for Related Time Series*. 12-13. U.S. Census Bureau, Center for Economic Studies.

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon D Woodcock (2009). *The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators*. In Producer Dynamics: New Evidence from Microdata, edited by Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. University of Chicago Press.

Department of Commerce, Office of Privacy and Open Government (2017). *Safeguarding Information*. http://osec.doc.gov/opog/privacy/pii_bii.html#PII

Drechsler, Jörg (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. New York: Springer.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006) Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third conference on Theory of Cryptography* (TCC'06), Shai Halevi and Tal Rabin (Eds.). Springer-Verlag, Berlin, Heidelberg, 265-284. DOI=http://dx.doi.org/10.1007/11681878_14

Haney, Samuel, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber (2017). *Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics*, SIGMOD'17, May 14-19, 2017, Chicago, Illinois, USA, DOI: 10.1145/3035918.3035940.

Karr, A.F., C.N. Kohnen, A. Oganian, J.P. Reiter, and A.P. Sanil (2006). *A framework for evaluating the utility of data altered to protect confidentiality*. The American Statistician 60, 224-232.

Kim, Hang J., Jerome P. Reiter, and Alan F. Karr (2016). *Simultaneously Edit-Imputation and Disclosure Limitation for Business Establishment Data*. Journal of Applied Statistics online: 1-20.

Lauger, Amy, Billy Wisniewski, and Laura McKenna (2014). *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research*. Research Report Series (Disclosure Avoidance #2014-02). Washington: Center for Disclosure Avoidance Research, U.S. Census Bureau.

Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber (2008). *Privacy: Theory Meets Practice on the Map*. Proceedings: International Conference on Data Engineering. Washington, DC, USA: IEEE Computer Society, 277-286.

Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk, and Peter Solenberger (2001). *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models*. Survey Methodology 27(1). Citeseer: 85-96.

U.S. Census Bureau (2016). OnTheMap: Data Overview (LODES Version 7). U.S. Census Bureau. https://lehd.ces.census.gov/doc/help/onthemap/OnTheMapDataOverview.pdf

Vilhuber, Lars and Ian M. Schmutte (2016). *Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy*. http://digitalcommons.ilr.cornell.edu/ldi/33/

Wang, Yue, Xintao Wu, and Donghui Hu (2016). *Using Randomized Response for Differentail Privacy Preserving Data Collection*. Workshop proceedings of the EDBT/ICDT 2016 Joint Conference. March 15, 2016, Bordeaux, France. http://ceur-ws.org/Vol-1558/paper35.pdf

9 Disclaimer

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

EXHIBIT 11

John M. Abowd et al., The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau (July 2020)

https://www2.census.gov/adrm/CED/Papers/CY20/2020-08-AbowdBenedettoGarfinkelDahletal-The%20modernization%20of.pdf

The modernization of statistical disclosure limitation at the U.S. Census Bureau

August 2020 (supersedes the 2017 version)

John M. Abowd[1], Gary L. Benedetto[2], Simson L. Garfinkel[3], Scot A. Dahl[4], Aref N. Dajani[2], Matthew Graham[5], Michael B. Hawes[2], Vishesh Karwa[6], Daniel Kifer[7], Hang Kim[8], Philip Leclerc[2], Ashwin Machanavajjhala[9], Jerome P. Reiter[10], Rolando Rodriguez[2], Ian M. Schmutte[11], William N. Sexton[12], Phyllis E. Singer[2], and Lars Vilhuber[2,12]

[1] Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau, John.Maron.Abowd@census.gov
[2] Center for Enterprise Dissemination, Disclosure Avoidance, U.S. Census Bureau, *firstname.m.lastname@census.gov*
[3] Senior Computer Scientist for Confidentiality and Data Access U.S. Census Bureau, Simson.L.Garfinkel@census.gov
[4] Economic Statistical Methods Division, U.S. Census Bureau, Scot.Alan.Dahl@census.gov
[5] Center for Economic Studies, U.S. Census Bureau, *firstname.m.lastname@census.gov*
[6] Department of Statistics, Harvard University, vkarwa@seas.harvard.edu
[7] Department of Computer Science and Engineering, Penn State University, dkifer@cse.psu.edu
[8] Department of Mathematical Sciences, University of Cincinnati, hang.kim@uc.edu
[9] Department of Computer Science, Duke University, ashwin@cs.duke.edu
[10] Department of Statistical Science, Duke University, jerry@stat.duke.edu
[11] Department of Economics, University of Georgia, schmutte@uga.edu
[12] Labor Dynamics Institute, Cornell University, {wns32,lv39}@cornell.edu

Abstract: Until recently, most U.S. Census Bureau data products used traditional statistical disclosure limitation (SDL) methods such as cell or item suppression, data swapping, input noise injection, and censoring to protect respondents' confidentiality. In response to developments in mathematics and computer science since 2003 that have significantly increased the risk of reconstruction and re-identification attacks, the Census Bureau is developing formally private SDL methods to protect its data products. These methods provide mathematically provable protection for respondent data and allow policy makers to manage the tradeoff between data accuracy and privacy protection—something previously done by technical staff. The first Census Bureau product to use formal methods for privacy protection was OnTheMap, a web-based mapping and reporting application that shows where workers are employed and where they live. Recent research for OnTheMap is implementing formal privacy guarantees for businesses to complement the existing formal protections for individuals. Research is underway to improve the disclosure limitation methods for the 2020 Census of Population and Housing, the American Community Survey, and the 2022 Economic Census. For each of these programs, we are developing new state-of-the-art privacy protection approaches based on formal mechanisms that have been vetted by the scientific community. There are many challenges in adopting formally private algorithms to datasets with high dimensionality and the attendant sparsity. In addition to formally private methods that allow senior executives to set the privacy-loss budget, our implementations will feature adjustable "sliders" for allocating the privacy-loss budget among related statistical products. The Census Bureau is implementing the settings for the privacy-loss budget and these sliders based on the decisions of the Census Bureau's Data Stewardship Executive Policy Committee.

1 Overview: Disclosure Limitation at the U.S. Census Bureau Today

The U.S. Census Bureau views disclosure limitation not just as a research interest, but as an operational imperative. The Census Bureau's hundreds of surveys and censuses of households, people, businesses, and establishments yield high quality data and derived statistics only if the Census Bureau maintains effective data stewardship and public trust.

The Census Bureau previously used traditional statistical disclosure limitation (SDL) techniques such as top- and bottom-coding, suppression, rounding, binning, noise injection, and sampling to preserve the confidentiality of respondent data. The Census Bureau is currently transitioning from these methods to modern SDL techniques based on formally private data publication mechanisms.

1.1 Legal Requirements

The Census Bureau collects confidential information from U.S. persons and businesses under the authority of Title 13 of the U.S. Code. Once collected, the confidentiality of that data is protected specifically by 13 USC §9, which prohibits:

(i) Using the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or
(ii) Making any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or
(iii) Permitting anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual records.

The privacy protections required by Title 13 are determined by the Census Bureau. Data users, including the Department of Justice and other government agencies, may be consulted regarding the criteria that determine fitness for use. Such consultation always respects the statistical-use-only requirement in the statute.

Some publications are further protected by Title 26 of the U.S. Code, which protects the federal tax information (FTI) used by the Census Bureau in the preparation of statistical products.

Confidentiality protection is intimately related to the statutory requirement that the published data be used for statistical purposes only. The definitions of "statistical purpose" and "nonstatistical purpose" were strengthened in Title III of the Foundations for Evidence-Based Policymaking Act of 2018, which is known as the Confidential Information Protection and Statistical Efficiency Act of 2018 (CIPSEA).

Additionally, the Department of Commerce (2017), in which the Census Bureau is housed, has issued directives regarding the protection of personally identifiable information (PII) and business identifiable information (BII). These directives largely mirror those issued by other government agencies and prohibit release of information

that can be used "to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc., alone or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."

1.2 Legacy methods supporting statistical disclosure limitation (SDL)

Historically, the Census Bureau has primarily used information reduction and data perturbation methods to support SDL (Lauger et al., 2014). Information reduction methods include top- and bottom-coding, suppression, rounding or binning, and sampling collected units for release in public use microdata files. Data perturbation methods include swapping, legacy noise injection systems, and partially and fully synthetic database construction. These legacy approaches start with the premise that there are specific data elements that must be protected (e.g., a person's income). A technical analyst chooses an approach from the assortment of available SDL methods that is likely to protect the data without resulting in too much damage to the published data accuracy. Usually, the selection of SDL method takes into consideration the intended uses of the published data along with assumptions about the kind of external data an intruder might have, and the types of privacy attacks an intruder might attempt.

These *ad hoc* approaches do not offer formal guarantees of data confidentiality. That is, there is no mechanism for quantifying how much privacy is being leaked from all publications based on a particular confidential database, or how one publication might interact with another publication or external data to create additional privacy risk. Furthermore, as the parameters of these legacy methods and their impact on the resulting accuracy of the data often needed to be kept confidential, there was limited opportunity for scientific scrutiny of their implementation or their effects.

1.3 Formal privacy approaches

Formal privacy methods take a different approach to protecting confidential information. Instead of starting with a list of confidential values to protect, an ad hoc collection of protection mechanisms, and ad hoc assumptions about attack models, the formal approach starts with a mathematical definition and framework for quantifying privacy risk, which permits the formulation of mathematically provable privacy guarantees against unwanted inference. Next, it implements mechanisms for publishing mathematical functions (typically called *queries*) based on the confidential data that are provably consistent with the formal privacy definition. Thus, data tables released by the statistical agency are actually modeled as a series of queries applied to the confidential data. Surrogates for public use microdata files can also be generated in this manner: instead of sampling the actual respondent data, queries are used to create formally private synthetic data. This is commonly done by first modeling the confidential data, then using the model to generate synthetic data, as discussed below.

Differential privacy (Dwork et al., 2006) is the most developed formal privacy method. It begins by specifying the structure of the confidential database to be protected, D. In

computer science, this is called the database schema; in statistics, it is referred to as the sample space. Two databases, D_1 and D_2, with the same schema are adjacent if the appropriately defined distance between them is, at most, unity. Leaving the technical details aside, say $|D_1 - D_2| \leq 1$. The universe of tables to be published from D is modeled as a set of queries on D, say Q. An element of Q, say q, is a single query on D. A randomized algorithm, A, takes as inputs D, q, and an independent random variable. The output of $A(D, q)$ is the statistic to be published, say S, which is a measureable set in the probability space defined by the independent random variable, say B. A randomized algorithm A for a publication system for releasing all of the queries in Q is ε-differentially private if, for all D_1 and D_2, with the same database schema and $|D_1 - D_2| \leq 1$, for all $q \in Q$, and for all $S \in B$:

$$\Pr[A(D_1, q) \in S] \leq e^{\varepsilon} Pr[A(D_2, q) \in S].$$

The probability is defined by the independent random variable that is used by the algorithm A, and not by the probability of observing any database D with the allowable schema (likelihood function in statistics).

There are alternative ways to define adjacent databases. For example, one method considers the databases adjacent if the record of a single person is added or removed from the database. Alternatively, the value of a single data item on a single record can be changed. Differential privacy is the mathematical formalization of the intuition that a person's privacy is protected if the statistical agency produces its outputs in a manner insensitive to the presence or absence of that person's data in the confidential database.

In differential privacy, the value ε is the measure of privacy loss or confidentiality protection. If $\varepsilon = 0$, then the two probability distributions in the definition always produce exactly the same answer from adjacent inputs—there is no difference in the output of algorithm A when given adjacent database inputs. Since the definition applies to the universe of potential inputs, and all databases adjacent to those inputs, all databases therefore produce exactly the same answer. Thus, the value $\varepsilon = 0$ guarantees no privacy loss at all (perfect confidentiality protection), but no data accuracy, since it is equivalent to releasing no data at all about the statistic S. In contrast, when $\varepsilon = \infty$, there is no confidentiality protection at all—full loss of privacy, but the statistic S is perfectly accurate (identical to what would be produced directly from the confidential input database). Thus, ε can be thought of as the *privacy-loss budget* for the publication of the queries in Q: the amount of privacy that individuals must give up in exchange for the accuracy that can be allowed in the statistical release.

Varying the privacy-loss budget allows us to move along a privacy-accuracy *Production Possibilities Frontier* (PPF) curve, as it is known in the economics literature, or along the *Receiver Operating Characteristics* (ROC) curve, as it is known in the statistics literature (Abowd and Schmutte 2019). For any attacker model, the curve constrains the aggregate disclosure risk that any confidential data might be jeopardized through any feasible reconstruction attack, given all published statistics. This budget is the worst-case limit to the inferential disclosure of any identity or item. In differential privacy,

that worst case is over all possible databases with the same schema for all individuals and items and over all external linking databases with any subset of that schema or those items.

The privacy-loss budget applies to the combination of *all* released statistics that are based on the confidential database. As a result, the formal privacy technique provides protection into the indefinite future and is not conditioned upon additional data that the attacker may have.

It is important to understand that the formal privacy protection offered by differential privacy is not absolute. Instead, it is a promise to individuals regarding the maximum amount of additional privacy loss that they may suffer as a result of a publication that is based in part on their confidential data.

To prove that a privacy-loss budget is respected, one must quantify the privacy-loss expenditure of each algorithm used to query the confidential data. The collection of the algorithms considered altogether must satisfy the privacy-loss budget. This means that the collection of algorithms used must have known composition properties.

Because the information environment is changing much faster today than when traditional SDL techniques were developed, it may no longer be reasonable to assert that a product is empirically safe given best-practice disclosure limitation prior to its release. Formal privacy models replace empirical disclosure risk assessment with designed protection. Resistance to all future attacks is a property of the design.

Differential privacy, the leading formal privacy method, is robust to background knowledge of the data, allows for sequential and parallel composability and for arbitrary post-processing edits, and enables full transparency of the implementation's source code. Differential privacy's proven guarantees hold even if external data sources are published or released later. Other formal privacy methods quantify the privacy loss that can also be mathematically established and proven, but with more constrained properties (e.g., Haney et al., 2017).

2 Expanding privacy protection for OnTheMap

Randomized response, a survey technique invented in the 1960s, was the first differentially private mechanism implemented by any statistical agency. Of course, randomized response was not recognized as being differentially private until *after* differential privacy was invented. Randomized response is sometimes called *local differential privacy.* Unfortunately, it is difficult to adapt randomized response to modern survey collection methods (Wang et al., 2016). It is the Census Bureau's experience that randomized response has a poor tradeoff between accuracy and privacy protection compared with the trusted curator model, and formal assessments of the expected additive errors of the two approaches confirm this (Kasiviswanathan et al., 2011). Vadhan notes "We have a better understanding of the local model than [multi-curator models where each trusted curator holds a portion of the confidential dataset.]

However, it still lags quite far behind our understanding of the single-curator model, for example, when we want to answer a set Q of queries (as opposed to a single query)." (Vadhan 2017)

The first production application of a formally private disclosure limitation system by any organization was the Census Bureau's OnTheMap (residential side only), a geographic query response system for studying residence and workplace patterns.

The Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES), the data used by OnTheMap, is a partially synthetic dataset that describes geographic patterns of jobs by their employment locations and residential locations as well as the connections between the two locations (U.S. Census Bureau, 2016). A job is counted if a worker is employed with positive earnings during the reference quarter and in the quarter prior to the reference quarter. These data and marginal summaries are tabulated by several categorical variables. The origin-destination (OD) matrix is made available by ten different "labor market segments". The area characteristics (AC) data–summary margins by residence block and workplace block–contain additional variables including age, earnings, and industry. The blocks are defined in terms of 2010 Census blocks, defined for the 2010 Census of Population and Housing. The input database is a linked employer-employee database, and statistics on the workplaces (Quarterly Workforce Indicators: QWI) are protected using noise injection together with primary suppression (Abowd et al., 2009, 2012).

For OnTheMap and the underlying LODES data, the protection of the residential addresses is independent of the protection of workplaces. Protection of worker information is achieved using a formal privacy model (Machanavajjhala et al., 2008); work is in progress to protect workplaces using formal privacy as well (Haney et al., 2017).

3 SDL methods supporting the 2020 Census of Population and Housing

The 2000 and 2010 Censuses of Population and Housing applied SDL in the form of record swapping, but this fact was not always obvious to data users. The actual swapping rate was kept confidential, as was the overall impact that swapping had on data accuracy (McKenna 2018).

The Census Bureau successfully tested the feasibility of producing differentially private tabulations of the redistricting data (PL94-171) for the 2018 End-to-End Census Test, and is currently in the final stages of algorithm development, for the full-scale implementation of differentially private protections for the 2020 Census of Population and Housing.

In October 2019 the Census Bureau re-released data from the 2010 Census using an early prototype for the 2020 Census Disclosure Avoidance System (DAS) (U.S. Census Bureau 2019). Called the 2010 Demonstration Data Products, this system was the subject of a December 2019 meeting of the Committee on National Statistics, where

attendees compared the statistical accuracy of these data products with previous data publications based on the 2010 Census. The source code used to prototype the 2010 Demonstration Data Products was released the following month. This code base included 33,853 lines of Python programs and 1263 lines of configuration files. In July 2020, the Census Bureau subsequently re-released the 2010 Census data protected using an updated version of the 2020 Census DAS, as the 2010 Demonstration Privacy-Protected Microdata File 2020-05-27 (U.S. Census Bureau 2020).

The differentially private mechanisms designed for the 2020 Census support the following products:

- **Public Law (PL) 94-171** files for redistricting;
- **Demographic Profiles and Demographic and Housing Characteristics files** for demographic statistics pertaining to individuals and housing units;
- **Detailed tabulations on race, ethnicity, and household composition;**
- **Privacy Protected Microdata,** the actual microdata from which published data products were tabulated; and
- **Noisy Measurements,** the actual differentially private statistics used to create the consistent microdata, to allow researchers outside the Census Bureau to produce independent statistical products without suffering the unavoidable accuracy loss that results from the post-processing of the differentially private statistics to convert them back into microdata for tabulation.

The Census Bureau has designed its differentially private algorithms to allow a selected number of queries based on the confidential data to be reported exactly. Such queries are called *invariants.* The Census Bureau currently plans the following invariants for the 2020 Census data publications:

- Total number of people by state;
- Total number of housing units (aggregate of occupied and vacant housing units) by block; and
- Total number of group quarters within three-digit group quarters type by block. Group quarters types are defined in Table P43 (U.S. Census Bureau 2012).[1]

While the inclusion of these invariants requires clarification of the formal privacy guarantees under differential privacy, they were considered necessary to permit public scrutiny of the state apportionment totals, and to permit the public-input component of the Local Update of Census Addresses (LUCA) program.

[1] Table P43, "Group Quarters Population by Sex and Age by Group Quarters Type," is in Segment 6 of the 2010 Census SF1. It can be downloaded from https://www2.census.gov/census_2010/04-Summary_File_1/.

Key disclosure limitation challenges include:

1. Ensuring consistency across tables by respecting the invariants enumerated above;
2. Producing block-level microdata for use by the Census Bureau's tabulation system to support production of traditional data products;
3. As was true of historical systems like swapping, there is difficulty detecting coding errors, particularly as they relate to verifying privacy-loss guarantees;
4. Determining how much of the privacy-loss budget should be spent per household; e.g., whether it should be proportional to household size;
5. A lack of high-quality usage data from which to infer relative importance of data products; and
6. The lack of public input data with which to develop and simulate the mechanism.

Key policy-related challenges include:

1. Communicating the global disclosure risk-data accuracy tradeoff effectively to the Data Stewardship Executive Policy Committee (DSEP) so that they can set the privacy-loss budget and the relative accuracy of different publications,
2. Providing effective summaries of the social benefits of privacy vs. data accuracy, so that DSEP, in particular, can understand how the public views these choices.

Throughout each decade, the Census Bureau also conducts special tabulations of small geographic areas such as towns. Those tabulations also impact privacy, and they also undergo SDL.

4 SDL methods supporting the American Community Survey (ACS)

The American Community Survey (ACS) is the successor to the long form survey of the Census of Population and Housing. The housing unit survey includes housing, household, and person-level demographic questions about a broad range of topics. There is a separate questionnaire for those residing in group quarters. The Census Bureau sends this survey to approximately 3.5 million housing units and group quarters each year and receives approximately 2.5 million responses. Weighted adjustments account for nonresponse, in-person interview subsampling, and controlling to pre-specified population totals. The ACS sample is usually selected at the tract level and is designed to allow reliable inferences for small geographic areas and for subpopulations, when cumulated across five years. ACS sampling rates vary across tracts. On average, a tract will have approximately thirty-five housing units and ninety people in the returned sample.

The Census Bureau releases one-year and five-year ACS data products. Five-year tables are released either by block group or by tract. One-year tables have been released only

for geographies containing at least 65,000 people. A recent Census Bureau Disclosure Review Board (DRB) decision allowed some one-year tables to be released for areas of at least 20,000, due to the termination of the three-year data products. The Census Bureau also releases one-year and five-year Public-Use Microdata Sample (PUMS) files for both persons and housing units. These PUMS contain samples of ACS microdata records (1% and 5% samples, respectively) with geographic detail limited to Public Use Microdata Areas (PUMA). PUMAs are special non-overlapping areas that partition each state into contiguous geographic units containing roughly 100,000 people.

The feasibility of developing formally private protection mechanisms given current methodological and computational constraints, the large number of ACS variables, and the desire for small area estimates is undemonstrated. The Census Bureau is actively pursuing this research, seeking to leverage advances from other data products. The Census Bureau is also funding cooperative agreement opportunities for research into the use of formal privacy for surveys in general. As an intermediate step to provide additional privacy to ACS respondents, the Census Bureau is experimenting with the development of non-formally private synthetic data using statistical and machine learning models to replace the current SDL methods.

Key disclosure avoidance challenges include:

1. **High dimensionality:** there are roughly two hundred topical module variables with mixed continuous and categorical values,
2. **Geography**, with estimates needed at the Census tract and block-group levels,
3. **Variable associations** across people in the same household,
4. **Outliers** in the economic variables,
5. **Survey weights** due to sampling, nonresponse, and population controls.

These challenges stem from high dimensionality combined with small sample sizes. Small geographies and sub-populations are important for data users, even if they do not always properly incorporate the sampling uncertainty when using these data. Tract-level and even block group-level data are critical for many applications, including the ballot language determinations in Section 203 of the Voting Rights Act. In addition to legislative districts, tabulations for many special geographies published by the Census Bureau, including cities and school districts, are built from smaller component geographies.

The large margins of error for small geographies allow some scope for introducing error from SDL without significantly increasing total survey error. Modelling can introduce some bias in exchange for massive decreases in variances by borrowing strength from correlations.

The research team is currently developing methods to protect ACS microdata utilizing synthesis models combined with a validation system. The overall approach is:

1. Build a chain of models, simulating each variable successively given the previous synthesized variables (Raghunathan et al., 2001). Currently, the team is assessing the use of classification trees for this purpose (Reiter, 2005);
2. Create synthetic microdata from these models for all records and all variables, creating fully synthetic data; and
3. Allow users to validate results from the synthetic microdata against the internal data. Validated results would have to meet the same standards for disclosure avoidance as all other public data releases and would be limited in quantity to statistics required for the stated purpose.

As opposed to current ACS Public Use Microdata Samples (PUMS), this fully synthetic microdata would not use internal files that have already had SDL applied to them as its source; rather, the ACS program will generate an Internal Reference File (IRF) to serve as the source. The IRF can serve as a baseline dataset for assessing survey accuracy without the confounding impacts of SDL methods, and will allow the research team to evaluate the effects of synthesis on privacy and accuracy in isolation.

The research team is considering other models for protecting tabular output, including hierarchical and spatio-temporal models.

Validation servers, verification servers[2], and access to the Federal Statistical Research Data Centers (FSRDCs) may be the solution for research questions for which the modernized SDL approach leads to reasonable uncertainty regarding the suitability of published data for a particular use. An advantage of the formally private methods being tested for both the 2020 Census and the ACS is that they permit quantification of the error contributed by the SDL; hence, the inferences drawn from these data can be corrected for the impact of the uncertainty added to protect privacy. Their suitability for use in a particular application can also be assessed without reference to the confidential data. This property of modernized SDL provides a means for applying objective criteria to a researcher's claim that the published data are suitable or unsuitable for a particular use.

5 SDL research supporting the 2022 Economic Census

Every five years the Census Bureau sends survey forms to nearly four million U.S. business establishments, broadly representative of all geographic regions and most private industries, to conduct the Economic Census. The Economic Census is based on a complete enumeration for certain types of businesses, and sampling of other, mostly smaller, businesses. The Census Bureau defines an *establishment* as a specific economic activity conducted at a specific location, and asks companies to file separate reports for

[2] Validation servers provide the data user with the results of their query calculated on the internal data with SDL performed on the result. Verification servers provide the data user with some measure of how confident they should be with the result of their query calculated on the synthetic data.

different locations and when multiple lines of activity are present at the same location. The Economic Census survey collects information from sampled establishments on the revenue obtained from product sales in the industries in which they operate, as well as information on employment, payroll, and other establishment characteristics.

Key policy challenges include:

1. Specifying the entity to be protected: multi-unit companies operate many establishments with different forms. From a legal standpoint, it is not entirely clear which entity (company, establishment, or something else) must be protected.
2. Defining what constitutes sufficient protection. Requirements to protect fact-of-filing may imply that whether a given business appears must be protected. However, it may not be necessary to protect certain business attributes that are in the public domain.

Key disclosure avoidance challenges include:

1. **Outliers** in the economic variables and generally high skewness;
2. **Sparsity** of data in cells disaggregated down to the North American Industry Classification System (NAICS) subsector and county level;
3. **Hybrid** sampling and enumeration design combined with an edit and imputation stage that complicate privacy models;
4. **Associations** among economic variables that increase disclosure risk; and
5. **Complex** publication schedules that require consistency over time and efficient allocation of privacy-loss budgets across releases.

The Census Bureau's disclosure modernization efforts for the Economic Census have followed two potentially complementary paths. Beginning in 2017, an interdisciplinary team at the Census Bureau partnered with academic colleagues to evaluate the feasibility of developing synthetic industry-level microdata. The methods under consideration are not formally private, but would allow publication of more detailed information while maintaining disclosure protections comparable to the cell suppression methods currently in use. Kim, Reiter, and Karr (2016) present methods of developing synthetic data on historic Economic Census data from the manufacturing sector. An inter-divisional team has applied two synthetic data models to 42 industries from the 2012 Economic Census covering eighteen economic sectors. Input data were limited to full-year reporter businesses (births, deaths, and seasonal businesses were excluded). The synthetic data were evaluated for fidelity in summary tabulations of items collected for all sectors. The team is currently evaluating the disclosure risk for these approaches. Kim and

Thompson are working on a separate synthetic data model that includes businesses that are part-year reporters.

In 2020 an additional team began work to develop formally private disclosure avoidance methods appropriate to economic data in general, and the Economic Census in particular. Since the publication schedule does not require release of microdata, the team is exploring modifications of the differential privacy paradigm that could be directly applied to tabular summaries and yield provable privacy guarantees. Specifically, they are considering a variant of the model developed in Haney et al., (2017) as well as other approaches in the smooth sensitivity framework (e.g. Nissim, Raskhodnikova and Smith, 2007). The sparsity of the published tables may require a modification of these methods to ensure consistency and data quality while keeping privacy loss at acceptable levels. The team intends to develop methods applicable to the County Business Patterns and Economic Census First Look products, which have relatively simple structure. From there it will hopefully be possible to adapt those methods to more complex Economic Census products.

6 Challenges and meetings those challenges

In differential privacy, the commonly used flattened histogram representation of the universe is calculated as the Cartesian product of all potential combinations of responses for all variables. This representation is often orders of magnitude larger than the total population even when structural zeroes (impossible combinations of values of variables, such as grandmothers who are three years of age) are imposed. One promising approach is approximate differential privacy, where the limiting factor depends only on the logarithm of the inverse probability of algorithmic failure.

Policy makers, including the Census Bureau's DSEP, must have enough information about the privacy-loss/data accuracy trade-off to make an informed decision about ε, and its allocation to different tabular summaries. In some cases, the chosen amount of noise injection from differential privacy may limit the suitability for use of the published statistics to more narrowly defined domains than has historically been the case.

The strategy for producing the tabular summaries is to supply the official tabulation software with formally private synthetic data that reproduce all of the protected tabulations specified in the redistricting and summary file requirements. In generating high quality synthetic microdata, one needs to consider integer counts, non-negativity, unprotected counts (e.g., total state population), and structural zeroes.

To execute this approach, the Census Bureau needs generic methods that will work on a broader range of datasets. In addition, it may be difficult to find meaningful correlations that are not represented in the model. To address this, the model must anticipate the types of analyses that data users might wish to conduct. As a result, better model-building tools are needed, as well as generic tools for correlating arbitrary models

with the ones used to build the synthetic data. Ongoing engagement with data users is also essential to help identify these intended uses of the published data.

Reproducible-science methods will be required to use synthetic data effectively.

Data are often collected with a complex sample design with considerable missing data and in panels of longitudinal data. Research is ongoing to ensure that weighted, longitudinal analysis using differentially private data will continue to produce "good results and good science" to the data users.

7 Approaches to gauge data accuracy and usefulness

There are multiple methods to assess data accuracy, also known as analytical (or inference) validity. Machanavajjhala et al. (2008) conducted experiments comparing differentially private synthetic data to the actual data for OnTheMap. They saw value in coarsening the domain to limit the number of "strange fictitious commuting patterns." Karr et al. (2006) and Drechsler (2011) advocate calculating confidence interval overlaps for parameters of interest, whether univariate, bivariate, or multivariate.

There is value in calculating all such metrics described above for parameter estimates calculated from:

- non-perturbed data (exact counts) where we expect parity; and
- parameter estimates that were not captured in the joint distributions modeled in the synthetic data, where one would not expect to uncover comparable results.

Disclosure limitation is a technology. It shows the relationship between privacy loss, which is considered a public "bad", and data accuracy, which is considered a public "good". A differentially private system can publish extremely disclosive data. This happens if the privacy-loss budget is set very high. The extremely disclosive data will likely be very accurate. That is, inferences based on these data will be nearly identical to those based on the confidential data. But extremely disclosive, albeit formally private, data also permit a very accurate reconstruction of the confidential data relative to the reconstruction possible with smaller privacy-loss budgets.

The teams at the Census Bureau working on formal privacy methods for statistical disclosure limitation have been charged by DSEP with developing technologies with adjustable parameters to control the privacy loss and data accuracy during implementation. Those technologies will be summarized with a variety of supporting materials. The Disclosure Review Board will make a recommendation regarding the appropriate formal privacy technology and parameter settings, including the privacy-loss parameter ε. The Data Stewardship Executive Policy Committee will review that recommendation and make the final determination. The published data will implement the recommendations of DSEP. Although more explicit than in previous censuses, this is the same chain of recommendation and approval that was used in 2000 and 2010.

This transition to innovation involves significant retooling of methods for the Census Bureau's career mathematical statisticians, computer scientists, subject matter experts, project and process managers, and internal stakeholders. This transition will help the Census Bureau lead similar innovation across the U.S. Federal Government and beyond.

8 References

Abowd, John M. and Ian M. Schmutte "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices," American Economic Review, Vol. 109, No. 1 (January 2019):171-202, DOI:10.1257/aer.20170627.

Abowd, John M., R. Kaj Gittings, Kevin L. McKinney, Bryce Stephens, Lars Vilhuber, and Simon D. Woodcock (2012). *Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures for Related Time Series.* 12-13. U.S. Census Bureau, Center for Economic Studies.

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon D Woodcock (2009). *The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators.* In Producer Dynamics: New Evidence from Microdata, edited by Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. University of Chicago Press.

Department of Commerce, Office of Privacy and Open Government (2017). *Safeguarding Information.* http://osec.doc.gov/opog/privacy/pii_bii.html#PII

Drechsler, Jörg (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation.* New York: Springer.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006) Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third conference on Theory of Cryptography* (TCC'06), Shai Halevi and Tal Rabin (Eds.). Springer-Verlag, Berlin, Heidelberg, 265-284. DOI=http://dx.doi.org/10.1007/11681878_14

Garfinkel, Simson, John M. Abowd, and Christian Martindale, Understanding Database Reconstruction Attacks on Public Data, Communications of the ACM, February 2019.

Garfinkel, Simson, John M. Abowd, Sarah Powazek, Issues Encountered Deploying Differential Privacy, Workshop on Privacy in the Electronic Society, Toronto, Canada - October 15, 2018.

Haney, Samuel, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber (2017). *Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics*, SIGMOD'17, May 14-19, 2017, Chicago, Illinois, USA, DOI: 10.1145/3035918.3035940.

Karr, A.F., C.N. Kohnen, A. Oganian, J.P. Reiter, and A.P. Sanil (2006). *A framework for evaluating the utility of data altered to protect confidentiality.* The American Statistician 60, 224-232.

Kasiviswanathan, Shiva Prasad, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith (2011). *What can we learn privately?.* SIAM Journal on Computing 40, no. 3: 793-826.

Kim, Hang J., Jerome P. Reiter, and Alan F. Karr (2016). *Simultaneously Edit-Imputation and Disclosure Limitation for Business Establishment Data.* Journal of Applied Statistics online: 1-20.

Lauger, Amy, Billy Wisniewski, and Laura McKenna (2014). *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research.* Research Report Series (Disclosure Avoidance #2014-02). Washington: Center for Disclosure Avoidance Research, U.S. Census Bureau.

McKenna, Laura (2018). *Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing.* Working Papers 18-47, Washington: Center for Economic Studies, U.S. Census Bureau.

Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber (2008). *Privacy: Theory Meets Practice on the Map.* Proceedings: International Conference on Data Engineering. Washington, DC, USA: IEEE Computer Society, 277-286.

Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk, and Peter Solenberger (2001). *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models.* Survey Methodology 27(1). Citeseer: 85-96.

U.S. Census Bureau (2012). 2010 Census Summary File 1: 2010 Census of Population of Housing. September 2012. U.S. Census Bureau. https://www.census.gov/prod/cen2010/doc/sf1.pdf

U.S. Census Bureau (2016). OnTheMap: Data Overview (LODES Version 7). U.S. Census Bureau. https://lehd.ces.census.gov/doc/help/onthemap/OnTheMapDataOverview.pdf

U.S. Census Bureau (2019). 2010 Demonstration Data Product. October 2019. U.S. Census Bureau. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html

U.S. Census Bureau (2020). 2010 Demonstration Privacy-Protected Microdata File 2020-05-27. July 2020. U.S. Census Bureau. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf/?#

Vadhan, Salil (2017). *The Complexity of Differential Privacy.* March 14, 2017. https://privacytools.seas.harvard.edu/files/privacytools/files/complexityprivacy_1_01.pdf

Vilhuber, Lars and Ian M. Schmutte (2016). *Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy.* http://digitalcommons.ilr.cornell.edu/ldi/33/

Wang, Yue, Xintao Wu, and Donghui Hu (2016). *Using Randomized Response for Differentail Privacy Preserving Data Collection.* Workshop proceedings of the EDBT/ICDT 2016 Joint Conference. March 15, 2016, Bordeaux, France. http://ceur-ws.org/Vol-1558/paper35.pdf

9 Disclaimer

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

EXHIBIT 12

U.S. Census Bureau, Decennial Census P.L. 94-171 Redistricting Data (Mar. 15, 2021)

https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html

Decennial Census P.L. 94-171 Redistricting Data

MARCH 15, 2021
CRVRDO

P.L. 94-171 Redistricting Data

Public Law (P.L.) 94-171, enacted by Congress in December 1975, requires the Census Bureau to provide states the opportunity to identify the small area geography for which they need data in order to conduct legislative redistricting. The law also requires the U.S. Census Bureau to furnish tabulations of population to each state, including for those small areas the states have identified, within one year of Census day.

Since the first Census Redistricting Data Program, conducted as part of the 1980 census, the U.S. Census Bureau has included summaries for the major race groups specified by the Statistical Programs and Standards Office of the U.S. Office of Management and Budget (OMB) in Directive 15 (as issued in 1977 and revised in 1997). Originally, the tabulation groups included White, Black, American Indian/Alaska Native, and Asian/Pacific Islander, plus "some other race." These race data were also cross-tabulated by Hispanic/Non-Hispanic origin. At the request of the state legislatures and the Department of Justice, for the 1990 Census Redistricting Data Program, voting age (18 years old and over) was added to the cross-tabulation of race and Hispanic origin. For the 2000 Census, these categories were revised to the current categories used today.

2020

In this section:

2020 Census P.L. 94-171 Redistricting Data Summary Files

Redistricting Data Summary Files expected by September 30, 2021

In declarations recently filed in the case of Ohio v. Raimondo, the U.S. Census Bureau made clear that we can provide a legacy format summary redistricting data file to all states by mid-to-late August 2021. Because we recognize that most states lack the capacity or resources to tabulate the data from these summary files on their own, we reaffirm our commitment to providing all states tabulated data in our user-friendly system by Sept. 30, 2021.

We have a prototype dataset in the legacy format available on our website. The data in this prototype is from our 2018 End-to-End Census Test in Providence, Rhode Island. However, it has the structure needed for understanding how this data will be published and for building a system to work with that data.

Declarations - Ohio v. Raimondo [https://www2.census.gov/programs-surveys/decennial/rdo/technical-documentation/2020Census/Combined_Declarations_Document.pdf] [<1 MB]

Prototype Redistricting Data [https://www.census.gov/programs-surveys/decennial-census/about/rdo/program-management.html#P3]

Technical Documentation

2020 Census State Redistricting (P.L. 94-171) Summary File Technical Documentation [https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/2020Census_PL94_171Redistricting_StatesTechDoc_English.pdf] [1.1 MB]

2020 Census State Redistricting (P.L. 94-171) Summary File Technical Documentation (Spanish)	[https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/2020Census_PL94_171Redistricting_StatesTechDoc_Spanish.pdf]	[1.1 MB]
2020 Census National Redistricting (P.L. 94-171) Summary File Technical Documentation	[https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/2020Census_PL94_171Redistricting_NationalTechDoc.pdf]	[1.1 MB]

The national documentation is only for the limited set of geographic entities which cross state boundaries.

Back to top [#top]

2020 Census P.L. 94-171 Geographic Support Products

TIGER\Line Shapefiles [https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html]

Use the 2020 Tab of the linked page.

Maps (.pdf format) [https://www.census.gov/geographies/reference-maps/2020/geo/2020pl-maps.html]

Block Assignment Files (BAFs) [https://www.census.gov/geographies/reference-files/time-series/geo/block-assignment-files.html]

Use the 2020 Tab of the linked page. BAFs are meant to be used in conjunction with the NLTs.

Name Look-up Tables (NLTs) [https://www.census.gov/geographies/reference-files/time-series/geo/name-lookup-tables.html]

Use the 2020 Tab of the linked page. NLTs are meant to be used in conjunction with the BAFs.

2010 to 2020 Tabulation Block Crosswalk Tables [https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.html]

Use the 2020 Tab of the linked page. Select Block Relationship Files.

Back to top [#top]

Group Quarters Assistance

The Census Bureau published a Federal Register Notice on the Final 2020 Census Residence Criteria and Residence Situations [https://www.federalregister.gov/documents/2018/02/08/2018-02370/final-2020-census-residence-criteria-and-residence-situations] on February 8, 2018. In the Notice, the Census Bureau stated we will continue the practice of counting prisoners at the correctional facility, to ensure that the concept of usual residence is interpreted and applied consistent with the intent of the Census Act of 1790. The Notice also stated the Census Bureau recognizes that some states have decided, or may decide in the future, to 'move' their group quarters (GQ) population (e.g. student, military, and prisoner population) to an alternate address for the purpose of redistricting. To assist those states, the Census Bureau is offering the use of a geocoding service. This service will be based on 2020 Census geographic data, by the end of February 2021, to assist states with their goals of reallocating GQ population for legislative redistricting.

November 04, 2019 | <P>CRVRDO </P>

Group Quarters Assistance - The Census Geocoder

[/programs-surveys/decennial-census/about/rdo/summary-files/2020/GQAssistance_CensusGeocoder.html]

Back to top [#top]

Additional 2020 Census Resources

Back to top [#top]

Related Information

Redistricting Data Program Management [/programs-surveys/decennial-census/about/rdo/program-management.html]

Redistricting Data Program [/rdo]

Decennial Census Datasets [/programs-surveys/decennial-census/data/datasets.html]

EXHIBIT 13

U.S. Census Bureau, Meeting Redistricting Data Requirements: Accuracy Targets (Apr. 7, 2021)

https://content.govdelivery.com/accounts/USCENSUS/bulletins/2cb745b

United States Census Bureau

Meeting Redistricting Data Requirements: Accuracy Targets

Meeting Redistricting Data Requirements: Accuracy Targets

Last year, the Census Bureau's Disclosure Avoidance System (DAS) Team made a number of important improvements to the TopDown Algorithm (TDA) that will be used to protect the privacy of our respondents' data in the P.L. 94-171 redistricting data product. As we have shown in our most recent set of demonstration data, those algorithmic improvements have substantially improved the accuracy of the resulting data, independent of the selection and allocation of the privacy-loss budget (PLB). As we explained in a recent newsletter, we have recently been turning our attention to optimizing and tuning the parameters of the algorithm to further improve accuracy.

The parameters of the TDA can be varied in a number of ways: query strategy, allocation of PLB across geographic levels, allocation of PLB across queries, and optimization of geographic post-processing to improve accuracy of the data for "off-spine" geographic entities. Determining the optimal settings for these parameters requires empirically evaluating large numbers of TDA runs against objective accuracy metrics.

Working with the Redistricting Community to Meet Data Requirements

For the P.L. 94-171 redistricting data product, the principal statutory use cases are the redistricting process and the U.S. Department of Justice's enforcement of the Voting Rights Act of 1965 (VRA). To facilitate this analysis, the Department of Justice supplied sample redistricting and VRA use cases for the Census Bureau to evaluate against.

Based on these use cases and additional feedback, we created an accuracy target to ensure that the largest racial or ethnic group in any geography entity with a total population of at least 500 people is accurate to within 5 percentage points of their enumerated value at least 95% of the time.

Because the redistricting and VRA use cases rely on geographic aggregations that cannot be prespecified (e.g., congressional districts that will be drawn after the data are published), for evaluation purposes the DAS Team used three already specified geographic constructs that resemble the size and composition of voting districts that will eventually be drawn: block groups (which are on the TDA geographic spine), census designated places (which are "off-spine"), and a customized set of off-spine entities that distinguished between strong minor civil division states and other states. The customized off-spine entities are similar to census designated places.

Because these accuracy targets are expressed in relative shares of the total population, tuning the TDA for accuracy of the racial/ethnic group's share also tunes the algorithm for corresponding accuracy of the total population of those geographies.

The Census Bureau is still evaluating the empirical results of these experimental runs, but we look forward to sharing these results and how they will inform the parameter settings used for our next set of demonstration data in our next newsletter.

2021 Key Dates, Redistricting (P.L. 94-171) Data Product:

By April 30:

- Census Bureau releases new Privacy-Protected Microdata Files (PPMFs) and Detailed Summary Metrics.
    - Two versions: Candidate strategy run using new PLB and old PLB.

By late May:

- Data users submit feedback.

Early June:

- The Census Bureau's Data Stewardship Executive Policy (DSEP) Committee makes final determination of PLB, system parameters based on data user feedback for P.L. 94-171.

Late June:

- Final DAS production run and quality control analysis begins for P.L. 94-171 data.

By late August:

- Release 2020 Census P.L. 94-171 data as Legacy Format Summary File*.

September:

- Census Bureau releases PPMFs and Detailed Summary Metrics from applying the production version of the DAS to the 2010 Census data.
- Census Bureau releases production code base for P.L. 94-171 redistricting summary data file and related technical papers.

By September 30:

- Release 2020 Census P.L. 94-171 data** and Differential Privacy Handbook.

* Released via Census Bureau FTP site.
** Released via data.census.gov.

Was this forwarded to you?

Sign up to receive your own copy!

Sign Up!

Useful Links:

- IPUMS NHGIS Privacy-Protected Census Demonstration Data
- DAS Updates
- Progress Metrics and Data Runs
- Newsletter Archives
- All DAS FAQs

Have Suggestions?

Do you have specific questions you'd like us to answer in this newsletter or topics you'd like discussed? Send us an email at 2020DAS@census.gov and let us know!

Contact Us

About Disclosure Avoidance Modernization

The Census Bureau is protecting 2020 Census data products with a powerful new cryptography-based disclosure avoidance system known as "differential privacy." We are committed to producing 2020 Census data products that are of the same high quality you've come to expect while protecting respondent confidentiality from emerging privacy threats in today's digital world.

Share This

Stay connected with us!
Join the conversation on social media.

EXHIBIT 14

Amy Lauger et al., U.S. Census Bureau, Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research 2 (Sept. 26, 2014)

https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/cdar2014-02-discl-avoid-techniques.pdf

RESEARCH REPORT SERIES
(Disclosure Avoidance #2014-02)

Disclosure Avoidance Techniques at the U.S. Census Bureau:

Current Practices and Research

Amy Lauger, Billy Wisniewski, and Laura McKenna

Center for Disclosure Avoidance Research
U.S. Census Bureau
Washington DC 20233

Report Issued: September 26, 2014

Abstract

The U.S. Census Bureau collects its survey and census data under the U.S. Code's Title 13, which promises confidentiality to its respondents. The agency also has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, the Census Bureau's goal is to release as much high quality data as possible while maintaining the pledge of confidentiality. We apply disclosure avoidance techniques prior to releasing our data products publicly to protect the confidentiality of our respondents and their data. This paper discusses the various types of data we release, the disclosure review process, restricted access procedures, disclosure avoidance techniques currently being used, and current disclosure avoidance research.

Key Words: Confidentiality, Microdata, Synthetic Data, Noise Infusion, Data Swapping

Table of Contents

1 Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This title prevents the Census Bureau from releasing any data "...whereby the data furnished by any particular establishment or individual under this title can be identified." In addition to Title 13, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires the protection of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. However, the agency certainly also has the responsibility and aim of releasing high quality data to the public for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality data as possible while maintaining the pledge of confidentiality. We apply disclosure avoidance techniques prior to releasing our data products publicly to protect the confidentiality of our respondents and their data. This paper discusses the various types of data we release, our disclosure review process, restricted access procedures, disclosure avoidance techniques currently used, and recent and current disclosure avoidance research. It is an update to Zayatz (2007).

2 Publicly Released Census Bureau Data

Unlike some statistical agencies, the Census Bureau does not use data licensing Massell and Zayatz, 2000) to provide data to some users but not to others. Therefore, all data released to any external party is considered publicly available. The Census Bureau uses different disclosure avoidance methods for each type of data before release to the public. The most common forms of data release are microdata, frequency count data, and magnitude data. The following sections will discuss the types of data we typically publish, the current methods we use to protect them, and recent and current research to improve our methods.

3 Microdata

3.1 Description

The Census Bureau releases microdata files from the decennial census, many demographic surveys, and some economic surveys. A microdata file consists of data at the respondent level. Each record represents one respondent and consists of values of characteristic variables for that respondent. Typical variables for a demographic microdata file are age, race, sex, income, and home ownership / tenure. Sometimes, files will focus on specific issues and might include variables about topics such as crime victimization and alcohol abuse.

Typically, the Census Bureau does not release microdata from economic surveys and censuses because the skewness of economic data makes it often easy to identify establishments by only a few characteristics. However, in recent years, the Census Bureau has produced a public use microdata file for the 2007 Survey of Business Owners and synthetic economic microdata files, such as the Survey of Income and Program Participation Synthetic Beta (SSB) and the synthetic Longitudinal Business Database (synLBD).

3.2 Current Disclosure Avoidance Methods

The Census Bureau currently uses several disclosure avoidance techniques for our microdata files including geographic thresholds, rounding, noise infusion, categorical thresholds, topcoding, and data swapping. This paper primarily describes the procedures used for the Census 2010 and American Community Survey Public Use Microdata Samples (PUMS) files but many of these techniques are also used for other microdata files. Of course, all direct identifiers (name, address, etc.) are removed before public release.

3.2.1 Geographic Thresholds

All geographic areas identified on public-use microdata files must have a population of at least 100,000 (Hawala, 2001). Several data sets have an even higher geographic threshold, which may, for example, only allow for the identification of the four Census Regions or the nine Census Divisions. Applicable thresholds are determined depending on the level of detail of the variables on the file, whether the survey is longitudinal, and the public availability of other similar data.

3.2.2 Rounding

The Census Bureau uses a traditional rounding scheme. For example, dollar amounts are rounded in this way:

$0 remains $0
$1-7 rounded to $4
$8-$999 rounded to nearest $10
$1,000-$49,999 rounded to nearest $100
$50,000+ rounded to nearest $1,000

Census 2000 data were used to develop this rounding scheme and the resulting rounded categories were deemed to have enough values in them. Rounding is done prior to all summaries and ratio calculations. Because the variable Property Taxes is readily and publicly available, it has larger categories than those resulting from the rounding described above. The variable Departure Time for Work is also rounded.

3.2.3 Noise Infusion

Sometimes, noise is added to demographic survey variables when other, more traditional protection methods are not suitable. For example, noise is added to the age variable for persons in households with 10 or more people. Ages are required to stay within certain groupings so certain statistics are not affected. The original ages are blanked and new ages are chosen from a given distribution of ages within their particular grouping. Noise is also added to a few other variables to protect small but well-defined populations but we do not disclose those procedures.

3.2.4 Categorical Thresholds

All categorical variables must have at least 10,000 people nationwide in each published category. Any categories not meeting this threshold must be recoded into broader intervals.

3.2.5 Topcoding

Topcoding is used to reduce the risk of identification by masking outliers in continuous variables. For example, someone with an income of five million dollars would appear to have a much lower income in the public data set. All continuous variables (age, income, travel time to work, etc.) are topcoded using the half-percent/three-percent rule. Topcodes for variables that apply to the total universe (e.g. age) must include at least 1/2 of 1 percent of all cases. For variables that apply to subpopulations (e.g. farm income), topcodes should include either 3 percent of the non-zero cases or 1/2 of 1 percent of all cases, whichever is the higher value. Distributions of data from the 1990 Census were used to develop this rule. Some variables, such as year born, are likewise bottomcoded.

3.2.6 Data Swapping

In data swapping, we identify "special uniques" (Elliott, et al. 1998), which are household records unique based on certain demographic variables at high levels of geography and thus have a substantial disclosure risk. Each such household is targeted to be swapped with another household in a different geographic area. Swapping typically does not affect many records. Swapping occurs at the microdata stage for the decennial census and for the American Community Survey but is performed primarily to protect aggregate data. See more about swapping in section 4.2.

3.3 Recent and Current Research

3.3.1 Re-identification Studies

The Census Bureau regularly conducts re-identification studies to assess the disclosure risk for our publicly available microdata. In light of the ever-changing amount, characteristics, and quality of other publicly available data, it is imperative for the Census Bureau to be situationally aware regarding the risk of our microdata products.

Most recently, Census Bureau staff and contractors conducted a re-identification study using public use microdata for the 2008 American Community Survey, other public information freely available on the Internet, and a demographic data set for three counties available for purchase. In the study, Census Bureau researchers found 926 unique records and successfully identified 87 people in these three counties. While this study shows that re-identification is fairly straightforward and possible, large-scale re-identification is not. Additionally, if an outsider intruder finds a possible match, it usually isn't a true match. Often survey records are unique within the sample but not in the population (Ramachandran, 2012). The Census Bureau will use the results of this research to continue to evaluate and adapt our disclosure avoidance procedures.

3.3.2 Synthetic Data

Given a data set, one can develop posterior predictive models to generate synthetic data that have many of the same statistical properties as the original data (Abowd and Woodcock, 2001). Synthetic data are often generated by sequential regression imputation one variable in one record at a time (Rubin, 1993). Using all of the original data, a regression model is developed for a given variable. Then, for each record, the original value of that variable is blanked and the model is used to impute a new value. In all, one follows these steps to create multiple synthetic populations. From here, one draws

random samples from the synthetic populations. These samples are the data that are released (Raghunathan, et al., 2003).

Synthesizing data can be done in different ways and for different types of data products. One can synthesize all variables for all records, known as a full synthesis, or one can synthesize a subset of variables for a subset of records, known as a partial synthesis. If doing partial synthesis, records that have a potential disclosure risk and those causing this risk are targeted. Generally, demographic data are modeled and synthesized more easily than economic data. Data can be synthesized with a goal of releasing the synthetic microdata or some other product generated from the synthetic microdata. Finally, one synthetic data set or implicate, which looks exactly like the original file, can be synthesized, or, alternatively, several different implicates could be released together. Multiple synthetic implicates can be analyzed using multiple imputation analysis techniques.

Through a partnership with Local Employment Dynamics (LED) partner states, the Census Bureau also releases a data product called OnTheMap. With version 6 as the latest release, OnTheMap is an online mapping and reporting tool that provides a user with information on where people are employed and where they reside, as well as connections between the two. Generally, data are available from all 50 states and U. S. territories from years 2002-2011, down to the Census block level. The underlying data come from a variety of sources, such as the LEHD Origin-Destination Employment Statistics (LODES), the Office of Personnel Management, and private workforce data from the Bureau of Labor Statistics.

OnTheMap is protected by strict confidentiality protection requirements. For example, residential address information for each workplace address is based on synthetic data, while workplace information is protected by some noise infusion. The Census Bureau is confident that the output does not disclose any confidential information.

A research group led by John Abowd of Cornell University has recently updated an existing public-use microdata file called the Survey of Income and Program Participation Synthetic Beta (SSB). This product links together individual-level microdata from the Census Bureau's Survey of Income and Program Participation, administrative tax data from the Internal Revenue Service, and retirement and disability benefit data from the Social Security Administration. Almost all variables on the file are synthesized, except for gender and the first marital link observed in the SIPP. The research group has determined that this new version cannot be linked to original SIPP public use files, nor SSB versions 4.0, 5.0 or 5.1 (Benedetto, et al, 2013). The Census Bureau approved the release of SSB 6.0 in June 2014.

The Synthetic Longitudinal Business Database (SynLBD) was the first business establishment-level public-use microdata file ever released by a U.S. statistical agency and was developed between researchers at Cornell University, Duke University, the National Institute of Statistical Standards (NISS), and the Census Bureau's Center for Economic Studies. (Jarmin, et al, 2014). This data set is fully synthetic, with all establishments and their characteristics modeled after the values in the confidential LBD. It contains information on 21 million establishment records across all sectors from 1976-2000. The current version does not include any geographic or firm-level variables.

4 Frequency Count Data

4.1 Description

The Census Bureau publishes frequency count data mainly from the decennial census and demographic surveys. Tables of frequency count data present the number of units in each table cell. For example, a table may have columns representing marital status and rows representing age groups. The cell values reflect the number of people in a given geographic area having the various combinations of marital status and age group. The decennial census and the American Community Survey have a multitude of published tables. However, other demographic surveys do not have a large enough sample to support tables at low levels of geography with sufficient data quality so only a limited number of tables at higher levels of geography are published.

4.2 Current Disclosure Avoidance Methods

Data swapping is the main procedure used to protect decennial census and American Community Survey tabulations. A small amount of household records is swapped with partner households in a different geographic area. The selection process to decide which households should be swapped is highly targeted to affect the records with the most disclosure risk. For example, households in very small geographic areas and those that are racially isolated are targeted. Households swapped with each other match on a minimal set of demographic variables. Public use microdata, tables, and all other data products are created from the swapped data files. After performing the data swapping for Census 2010, the Census Bureau did an extensive evaluation of the procedure and the resulting tables' preservation of data quality. The results of this evaluation are confidential but the effects of the data swapping were minimal compared to sampling, measurement, coverage, and non-response error already present.

The Census Bureau continually conducts research to adapt and improve the swapping procedures. Over the past few years, we have altered the swapping routine, changed the variables used to determine which households are at risk, and slightly increased the percentage of households that are swapped.

Synthetic data are used to protect some of the data from the decennial census and the American Community Survey. Both programs collect data for both residential households and group quarters. Swapping is infeasible for group quarters so we now use partially synthesized group quarters data for these programs (Hawala, 2008). The Census Transportation Planning Products (CTPP) special tabulations also use synthetic data (Li, et al., 2011).

Tables are often required to meet certain thresholds in order to be released. For example, Summary File 2 for the decennial census iterates a set of tables by universe groups such as race, ancestry, and ethnicity. For these tables, each universe must contain at least 100 people in a given geographic area to be released. The American Community Survey has several types of rules, including population thresholds and geographical restrictions, some for data quality for its 1- and 3-year data products and some for disclosure avoidance (U.S. Census Bureau, 2013).

Often the standard products for the decennial census and the American Community Survey do not include the data particular users need. These users can request and pay for a special tabulation. All

special tabulations are generated from the swapped data files and must meet certain criteria before release.

All cell values are rounded according to the following scheme:

0 remains 0
1-7 rounds to 4
8 or greater rounds to the nearest multiple of 5

Totals are constructed before rounding, so the universes remain the same from table to table but the tables may no longer be additive. Percentages and rates are calculated after rounding. We allow some exceptions when the numerator, denominator, or both are not shown.

Tables usually must have no more than three or four dimensions and a mean cell size of at least three and sometimes higher than that. Thresholds on universes are often applied to avoid showing data for small geographic areas or small population groups. Usually any cells with an unweighted count of one or two are not published and, for survey data, usually only weighted estimates are published.

Percentiles and other quantiles may be calculated in one of two ways. If they are calculated as an interpolation from a frequency distribution of unrounded data, no additional rounding is required. Otherwise, they must be rounded to two significant digits and at least five observations must be on either side of each quantile point.

4.3 Recent and Current Research

The Census Bureau continues to research ways to improve protection of frequency count data. Recent research explored two methods to improve data swapping. The research involved two new aspects. The first method is the use of "n-cycles" for swapping instead of swapping pairs of households with each other. In the current method, one could say the Census Bureau uses a swap cycle of size two, with two households, say A and B. Household A's characteristics are swapped with the characteristics of household B. In the n-cycle approach, the cycle may involve more than two households. For example, if n=3, A's characteristics are assigned to B, B's characteristics are assigned to C, and C's characteristics are assigned to A. Unlike the current method, in the case of an odd number of households for a given set, the new method will allow all households to be swapped. The second explored method for swapping involved the creation of a method to rank swaps in terms of data utility versus disclosure risk (DePersio, et al, 2012). The results were favorable but are not yet implemented into Census Bureau data products.

Additionally, researchers are currently studying the use of post-randomization (PRAM) methods as an alternative to data swapping, with a paper forthcoming.

5 Magnitude Data

5.1 Description

The Census Bureau publishes magnitude data from many of its surveys and the economic census. Most magnitude data comes from economic data products. However, some demographic variables such as household income is in the form of magnitude data. For economic data, tables of magnitude data usually contain both the frequency counts of establishments in each cell and the aggregate of some quantity of interest over all units (e.g., establishments) in each cell. For example, a table may present the total value of shipments within the manufacturing sector by North American Industry Classification System (NAICS) code by county. The frequency counts in the tables are not considered sensitive because so much information about establishments, particularly classifications that would be used in frequency count tables, is publicly available. However, the magnitude values are considered sensitive and must be protected. Magnitude data are generally non-negative quantities. A given firm may have establishments that are in more than one table cell. Protection is applied to the firm level rather than the establishment level. Disclosure avoidance techniques are used to ensure published data cannot be used to estimate an individual firm's data too closely.

5.2 Current Disclosure Avoidance Methods

5.2.1 Cell Suppression

The Census Bureau uses cell suppression for disclosure avoidance for most of its tables of magnitude data in economic data products. Any table cell value that could allow users to estimate a responding company's value too closely is not shown. The value is suppressed and replaced with a "D" for disclosure. These sensitive cells are called primary suppressions. They are identified using the p% rule, which is designed to ensure that a user cannot estimate a respondent's value to within p% of that value (Federal Committee on Statistical Methodology, 20054).

Because marginal totals are shown in the tables, other cells called complementary suppressions must be selected and suppressed, so that primary suppression values cannot be derived or estimated too closely via addition and subtraction of published values. For the past few years, researchers have worked on developing new cell suppression software. The modernized software is based on linear programming and replaces the older system that relied on network flow theory.

The new system is able to protect certain classes of tables better than the old system. Significantly, linear programming now allows for precise protection of three-dimension tables, as well as most sets of linked tables. The Census Bureau is required to protect economic data at the firm level, as well as at the establishment level. In order to improve on this requirement, the system implements a new feature, called "protection of supercells." Here, a supercell is defined as the union of all interior primaries, along with the set of all secondaries, which exist in specified additive constraints (Massell, 2011). In addition, linear programming eliminates under-suppression and reduces over-suppression. Thus, more data can be published while still fulfilling protection requirements. The new system includes several innovative algorithmic procedures that allow the program to run quickly enough to meet production requirements (Steel, 2013).

5.2.2 Noise Infusion

A different technique is used for many of the Census Bureau's economic data products. This technique, commonly referred to as EZS noise, is applied to the underlying microdata prior to tabulation (Evans, et al, 1998). Each responding company's data are perturbed by a small amount, say up to 10% in either direction. The actual percentage used by the Census Bureau is confidential. Noise is added in such a way that cell values that would normally be primary suppressions, thus needing protection, are changed by a large amount, while cell values that are not sensitive are changed by a small amount. Noise has several advantages over cell suppression – it enables data to be shown in all cells in all tables, it eliminates the need to coordinate cell suppression patterns between tables, and it is a much less complicated and less time-consuming procedure. Because noise is added at the microdata level, additivity of the table is guaranteed.

To perturb an establishment's data by about 10%, the Census Bureau multiplies its data by a random number that is close to either 1.1 or 0.9. Any of several types of distributions may be used from which to choose our multipliers and the distributions remain confidential within the agency. The overall distribution of the multipliers is symmetric about 1. The noise procedure does not introduce any bias into the cell values for census or survey data. Because we protect the data at the firm level, all establishments within a given firm are perturbed in the same direction. The introduction of noise causes the variance of an estimate to increase by an amount equal to the square of the difference between the original cell value and the noise-added value. One could incorporate this information into published coefficients of variation.

The following surveys now use noise infusion to protect their data: Nonemployer Statistics, Integrated Longitudinal Database, the LEHD Quarterly Workforce Indicators, workplace information for OnTheMap, Commodity Flow Survey, Survey of Business Owners, and County Business Patterns. Cell suppression is still the method of choice for the stateside Economic Census but noise infusion is now used for the Economic Census of Island Areas.

In some surveys whose data are protected using noise, a single table is considered to be the most important one. For these surveys, staff developed an enhanced version of the EZS methodology, called "balanced noise." Here, noise factors are not assigned randomly to each of the microdata records. Instead, select records are placed into small groups, which are defined by the unique interior cells of the table to which they contribute. The noise factors are then assigned to each of these groups by alternating the direction of the noise factors to each contributing record. This process enhances the amount of noise cancellation in most cells and results in cells closer to the true values. Balanced noise is more complicated to implement than random EZS noise but the improved accuracy of the "most important table" is often worth the extra effort. Massell and Funk found that the effect of balanced noise on one table does not typically hurt the accuracy on other produced tables, while guaranteeing the protection of the underlying microdata (2007).

5.2.3 Synthetic Data

Many external users are interested in having the Census Bureau release more microdata from its surveys and censuses. However, releasing microdata poses many risks due to the great amount of data readily

available on the Internet. Currently, the following economic data products use synthetic data to protect the underling data: OnTheMap versions 3-6, SIPP Synthetic Beta (SSB), and the Synthetic Longitudinal Business Database (SynLBD). The SSB and the SynLBD are available through the Cornell University Virtual RDC.

5.3 Recent and Current Research

Recall that in cell suppression, the Census Bureau uses the p% rule to identify sensitive cells. This rule is designed to ensure that a user cannot estimate a respondent's value to within p% of that value. Currently, staff use fixed interval protection, which means the lower bound of the interval of uncertainty around any respondent's value *v* must be at most (1-p/100)* *v* and the upper bound must be at least (1+p/100) * *v*. This rule ensures that both bounds are a given distance from the true value.

The Census Bureau is currently developing a tabular statistical disclosure control method that combines some of the best features of cell suppression, noise addition, and rounding. The resulting table would have no suppressed cells but each value would have an uncertainty associated with it. This uncertainty would be expressed as the value plus or minus an error term.

Another current focus is about applying the p% rule to atypical types of data, such as percentages, rounded data, negative values, differences, net changes, and weighted averages.

6 The Disclosure Review Board

The Census Bureau has a Disclosure Review Board (DRB), which establishes disclosure avoidance policies and ensures consistency in the disclosure review of all publicly released Census Bureau data products. The board consists of at least six members representing the Census Bureau's demographic, decennial, and economic directorates, and the Research Data Centers (RDCs). These members usually serve six-year terms. At least an additional three members representing the research and policy areas are permanent members.

The Disclosure Review Board reviews almost all publicly released data products as explained in the DRB checklist (U. S. Census Bureau, 2007). These data products include those produced by Census Bureau staff and those produced at the Research Data Centers. Census Bureau staff members wishing to release data send a memo to the chair of the DRB accompanied by the DRB checklist, the questionnaire from the survey or census, a list of variables of interest, a record layout for requested microdata, table outlines for requested tabular data, and often some cross-tabulations of the variables of interest. The DRB checklist asks basic questions about the content of the data file to be released and helps to ensure consistency in the DRB's decision-making process. The Federal Committee on Statistical Methodology has created a generalized checklist (1999) for use by other federal statistical agencies.

After reviewing a request, the DRB may choose to approve it as is, approve it with modifications, or deny it. Census Bureau staff members not satisfied with a decision may appeal the decision to the Data Stewardship Executive Policy Committee (DSEP), which consists of a subset of Census Bureau Associate Directors.

7 Research Data Centers

Some data sets cannot be publicly released because of confidentiality concerns. However, we have developed some restricted-use data procedures to allow researchers to use Census Bureau data in a secure environment at what is known as Research Data Centers (RDCs). To use the RDCs, researchers must submit a proposal to the Census Bureau stating what research they wish to conduct, which restricted data sets they will need, and what type of results are to be published. The research must benefit the Census Bureau in some way, such as by improving data quality or improving methodology to collect, measure, or tabulate a survey, census, or estimate. If the proposal is accepted, the researcher and any associates who will work on the project at the RDC must obtain Special Sworn Status and come to one of the RDCs to work with the data they need. The researchers are then required by law to maintain confidentiality for life, just as any other Census Bureau employee is. Census Bureau staff review research results for disclosure problems before they are publicly released. Currently, eighteen RDCs span the country with more opening often.

8 Conclusion

Several developments have occurred in disclosure avoidance methodology at the Census Bureau since Zayatz (2007) was published. The noise infusion technique for establishment magnitude data is used in more economic data sets. Improved data swapping techniques have been performed on Census 2010 and American Community Survey data and research continues on ways to improve the technique further. Re-identification experiments on our microdata files continue. Current research focuses on synthetic data, the Microdata Analysis System, and other new disclosure avoidance alternatives for both demographic and economic data.

9 References

Abowd, J. and S. Woodcock. 2001. "Disclosure Limitation in Longitudinal Linked Data." *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* Edited by P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes. 215-277. Netherlands: Elsevier Science.

Benedetto, G., M. Stinson, and J. Abowd 2013. "The Creation and Use of the SIPP Synthetic Beta." Suitland, MD: U.S. Census Bureau. Available at: http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf (accessed August 2014).

DePersio, M., K. Ramanayake, J. Tsay, L. Zayatz 2012. "n-Cycle Swapping for the American Community Survey." In *Privacy in Statistical Databases 2010 LNCS 7556*. Edited by J. Domingo-Ferrer and I. Tinnirello. 143-164. Berlin: Springer Verlag.

Elliott, M., C. Skinner, and A. Dale. 1998. "Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk." *Research in Official Statistics* 1: 53-67. Luxembourg, March 1998.

Evans, B., L. Zayatz, and J. Slanta 1998. "Using Noise for Disclosure Limitation for Establishment Tabular Data." *Journal of Official* Statistics 14 no 4: 537-551. Available at http://www.jos.nu/Articles/abstract.asp?article=144537 (accessed August 2014).

Federal Committee on Statistical Methodology 1999. "Checklist on Disclosure Potential of Proposed Data Releases." Washington, DC: U. S. Office of Management and Budget. Available at http://fcsm.sites.usa.gov/files/2014/04/checklist_799.doc (accessed August 2014).

Federal Committee on Statistical Methodology 2005. *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology* Version 2. Washington, DC: U. S. Office of Management and Budget. Available at http://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf (accessed August 2014).

Jarmin, R., T. Louis, and J. Miranda. 2014. "Expanding the Role of Synthetic Data at the US Census Bureau." Statistical Journal of the IAOS: Journal of the International Association for Official Statistics 30 no 2: 117-121. DOI: http://dx.doi.org/10.3233/SJI-140813.

Hawala, S. 2001. "Enhancing the "100,000 Rule: On the Variation Of The Per Cent Of Uniques in a Microdata Sample And the Geographic Area Size Identified On The File." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 2001. Available at http://www.amstat.org/sections/srms/proceedings/y2001/Proceed/00211.pdf (accessed August 2014).

Hawala, S. 2008. "Producing Partially Synthetic Data to Avoid Disclosure." In Proceedings of the Section on Government Statistics: American Statistical Association, August 2008. 1345-1350. Alexandria, VA. American Statistical Association. Available at https://www.amstat.org/sections/srms/Proceedings/y2008/Files/301018.pdf (accessed August 2014).

Li, J., T. Krenzke, M. Brick, D. Judkins, M. Larsen. 2011. "Variance Estimation for the Census Transportation Planning Products with Perturbed American Community Survey Data." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 2011. 1595-1603. Alexandria, VA. American Statistical Association. Available at https://www.amstat.org/sections/srms/proceedings/y2011/Files/301081_66127.pdf (accessed September 2014).

Massell, P. and L. Zayatz. 2000. "Data Licensing Agreements at U.S. Government Agencies and Research Organizations." In Proceedings of the International Conference on Establishment Surveys II. July 2000. 1393-1410. Alexandria, VA. American Statistical Association. Available at http://www.amstat.org/meetings/ices/2000/proceedings/S29.pdf (accessed August 2014).

Massell, P. and J. Funk, 2007. "Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection." In Proceedings of the 2007 Federal Committee on Statistical Methodology (FCSM) Research Conference. Available at http://fcsm.sites.usa.gov/files/2014/05/2007FCSM_Massell-IX-B.pdf (accessed September 2014).

Massell, P. 2011. "Modernizing Cell Suppression Software at the U. S. Census Bureau." Proceedings of the Section on Survey Research Methods, American Statistical Association, August 2011. 3007-3015. Alexandria, VA. American Statistical Association. Available at http://www.amstat.org/sections/srms/proceedings/y2011/Files/301855_67474.pdf (accessed August 2014).

Raghunathan, T., J. Reiter, and D. Rubin 2003. "Multiple Imputation for Statistical Disclosure Limitation." Journal of Official Statistics 19 no 1: 1-16. Available at http://www.jos.nu/Articles/abstract.asp?article=191001 (accessed August 2014).

Ramachandran, A., L. Singh, E. Porter, and F. Nagle. "Exploring Re-Identification Risks in Public Domains." 2012 Tenth Annual International Conference on Privacy, Security and Trust (July 2012). Danvers, MA: Institute of Electrical and Electronics Engineers. DOI:10.1109/pst.2012.6297917.

Rubin, D. 1993. "Discussion of Statistical Disclosure Limitation." Journal of Official Statistics 9 no 2: 461-468. Available at http://www.jos.nu/Articles/abstract.asp?article=92461 (accessed September 2014).

Steel, P. 2013. "The Census Bureau's New Cell Suppression System." In Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference. https://fcsm.sites.usa.gov/files/2014/05/E3_Steel_2013FCSM.pdf (accessed August 2014).

Zayatz, L. 2007. "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. Journal of Official Statistics 23 no. 2: 253-265. Available at http://www.jos.nu/Articles/abstract.asp?article=232253 (accessed August 2014.)

U.S. Census Bureau 2007. "Supporting Document Checklist on Disclosure Potential of Data Disclosure Review." Suitland, MD: U.S. Census Bureau. Available at https://www.census.gov/srd/sdc/S14-1_v1.3_Checklist.doc (accessed August 2014).

U.S. Census Bureau 2013. “American Community Survey: Data Suppression.” Suitland, MD: U. S. Census Bureau. Available at http://www.census.gov/acs/www/Downloads/data_documentation/data_suppression/ACSO_Data_Suppression.pdf (accessed August 2014).

EXHIBIT 15

JASON, Formal Privacy Methods for the 2020 Census (Apr. 2020)

https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/privacy-methods-2020-census.html

Formal Privacy Methods for the 2020 Census

Contact: Gordon Long — glong@mitre.org

April 2020

JSR-19-2F

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 983-6997

Contents

Abstract

In preparation for the 2020 decennial census, the Census Bureau asked JASON to examine the scientific validity of the vulnerability that the Census Bureau discovered in its traditional approach to Disclosure Avoidance, the methods used to protect the confidentiality of respondent data. To address this vulnerability, the Census Bureau will employ differential privacy, a mathematically rigorous formal approach to managing disclosure risk. JASON judges that the analysis of the vulnerability performed by Census is scientifically valid. The use of Differential Privacy in protecting respondent data leads to the need to balance statistical accuracy with privacy loss. JASON discusses this trade-off and provides suggestions for its management.

1 EXECUTIVE SUMMARY

A decennial population census of the United States will officially begin April 1, 2020. Under Title 13 of the US Code, the Bureau of the Census is legally obligated to protect the confidentiality of all establishments and individuals who participate in providing census data. In particular, Census cannot publish any information that could be used to identify a participant.

Over the years, a large amount of personal data have become easily available via online and commercial resources. It has also become much easier to analyze large amounts of data using modern computers and data-science tools. This has made it possible to breach the confidentiality protection promised to respondents of studies and surveys. There have been several notable examples in which records collected under pledges of confidentiality from a survey were linked with public data resulting in the re-identification of the individuals participating in the survey. In an exercise to evaluate the confidentiality protection of the census, the Census Bureau discovered such a vulnerability exists for their data as well.

Using the individual responses from participants (known as microdata), the Census Bureau produces a collection of tables that summarize population counts, age distributions, etc., for various levels of geographic resolution from the whole nation down to census blocks. A variety of approaches have been used by Census in the past to prevent re-identification. In addition to the removal of direct identifiers, Census applies geographic thresholding, top and bottom coding, swapping and other methods of obfuscation to hide identifying characteristics. It was previously thought to be computationally intractable to reconstruct the microdata from the published tabular summaries. But in 2018, applying modern optimization methods along with relatively modest computational resources, Census succeeded in reconstructing, from the published 2010 census data, geographic location (census block), sex, age, and ethnicity for 46% of the US population (142 million people). By linking the reconstructed microdata with information in commercial

databases, Census was then able to match and putatively re-identify 45% of the reconstructed records. Of these putative re-identifications, 38% were confirmed. This corresponds to 17% of the US population in 2010 (a total of over 52 million people). Such a re-identification rate exceeds that obtained in a previous internal Census assessment by four orders of magnitude. Public release of these re-identifications would constitute a substantial abrogation of the Census' Title 13 confidentiality obligations.

In view of these developments, Census has proposed the application of formal privacy methods, in particular, the use of Differential Privacy (DP). DP, introduced in 2006, has as its goal the prevention of learning about the participation of an individual in a survey by adding tailored noise to the result of any query on data associated with that survey. DP provides a set of algorithms used to compute statistical information about a dataset (e.g. counts, histograms, etc.), but infuses those statistics with tailored noise, making it possible to publish information about a survey while limiting the possibility of disclosure of detailed private information about survey participants.

A number of features make DP an attractive approach for protection of confidentiality for the 2020 census and beyond. Notably, privacy loss (in a technical sense) can be rigorously quantified via a privacy-loss parameter. In addition, there are techniques to create synthetic data such that subsequent queries will not cause further confidentiality loss provided such queries do not access the original data. Finally, confidentiality would degrade in a controlled way should it prove necessary to re-access the original data in order to publish further tabulations. Census proposes to use this approach by adding noise to the tabular summaries it traditionally publishes and then using these to reconstruct synthetic census microdata. Both the noised tabular summaries and the synthetic microdata could then be publicly released.

Once the differentially private tabulations and the synthetic data are produced, the use of DP methods offers a mathematically rigorous guarantee that any

further analysis of the released data preserves the original level of confidentiality protection. However, one drawback of such approaches is that the applied noise will degrade the accuracy of various tabulations and statistical analyses of the data, particularly those associated with small populations. Census data are used by a large number of government, academic, business, and other stakeholders. Census is therefore compelled to make an explicit trade-off between the accuracy of its data releases and the privacy of respondents.

Census charged JASON with the following three tasks:

1. Examine the scientific validity of the vulnerability that the Census Bureau discovered in the methods that it has historically used to protect the confidentiality of respondent data when preparing publications;
2. Evaluate whether the Census Bureau has properly assessed the vulnerability as described above;
3. Provide suggestions to represent the trade-offs between privacy-loss and accuracy to explicitly represent user choices.

JASON has not attempted to duplicate the reconstruction of census microdata as it does not have access to that data, nor to data from commercial marketing databases. JASON has, however, confirmed via database simulation that such an attack is possible; JASON concludes that, provided one publishes a sufficient number of tabular summaries, there are multiple approaches using modern optimization algorithms to reconstruct the database from which the summaries originated with high probability. This creates a significant risk of disclosure of census data protected under Title 13.

Census plans to release some data without noise, most importantly, state populations for the apportionment of Congressional representatives. In addition, Public Law 94-171 requires that Census provide the states with small-area data necessary to perform legislative redistricting for both Federal and State electoral

districts. The Census has set up a voluntary program in which state officials define the geographic areas for which they wish to receive census data. While only population data are legally mandated, Census has traditionally also provided other demographic data such as race, ethnicity and voting age populations. For expedience, states have simply asked for these data at the finest geographical resolution (census blocks) and have then used the block populations to infer population counts for larger geographical areas such as legislative districts. The proposed creation of differentially private census tabulations will result in block-level populations that differ from the original census enumeration due to the infused noise. Releases of exact counts (known as invariants) are technically violations of DP in principle and degrade the privacy guarantee, although to what extent in practice remains a research issue. There arises, then, a tension between the obligations under PL 94-171 to release population data for legislative purposes and the requirements of Title 13 to protect confidentiality.

For large populations, for example at the national, state, or even in many cases the county level, using DP does not unduly perturb the accuracy of statistical queries on the data provided the privacy-loss parameter is not set too low (implying the infusion of a large amount of noise). This is important for diverse users of census data (demographers, city planners, businesses, social scientists etc.). But as the size of the population under consideration becomes smaller, the contributions from injected noise will more strongly affect such queries. Note that this is precisely what one wants for confidentiality protection, but is not desirable for computation of statistics for small populations. Thus there is also a tension between the need to protect confidentiality and the aim to provide quality statistical data to stakeholders. While the latter is not legally mandated for Census, it is aligned with the Office of Management and Budget's policy directive to agencies that produce useful governmental statistics, and Census has traditionally been a key supplier of such data through its various published products.

The trade-off between confidentiality and statistical accuracy is reflected in the choice of the DP privacy-loss parameter. A low value increases the level of

injected noise (and thus also confidentiality) but degrades statistical calculations. Another factor that also influences the choice of privacy-loss parameter is the number and geographical resolution of the tables released. For example, if no block-level data were publicly released, a re-identification "attack" of the sort described above presumably would become more difficult, perhaps making it feasible to add less noise and thus publish tables at a higher value of the privacy loss parameter than what would be required if block level tables were published. A re-identification attack, of the sort that originally led to the conclusions that more rigorous and effective confidentiality protections were required, has not been performed on microdata reconstructed from differentially private tabulations. Such an analysis is needed to gauge the level of protection needed.

Depending on the ultimate level of privacy protection that is applied for the 2020 census, some stakeholders may well need access to more accurate data. A benefit of differential privacy is that products can be released at various levels of protection depending on the level of statistical accuracy. The privacy-loss parameter can be viewed as a type of adjustable knob by which higher settings lead to less protection but more accuracy. However, products publicly released with too low a level of protection will again raise the risk of re-identification. One approach is to use technology (e.g. virtual machines, secure computation platforms etc.) to provide protected data enclaves that allow access to census data at lower levels of privacy protection to trusted stakeholders. Inappropriate disclosure of such data could still be legally enjoined via the use of binding non-disclosure agreements such as those currently in Title 13. This idea is similar to the concept of "need to know" used in environments handling classified information.

Finally, it will be necessary to engage and educate the various communities of stakeholders so that they can fully understand the implications (and the need for) DP. These engagements should be two-way conversations so that the Census Bureau can understand the breadth of requirements for census data, and stakeholders can in turn more fully appreciate the need for confidentiality protection in the present era of "big data", and perhaps also be reassured that their statistical

needs can still be met.

1.1 Findings

1.1.1 The re-identification vulnerability

- The Census has demonstrated the re-identification of individuals using the published 2010 census tables.
- Approaches to disclosure avoidance such as swapping and top and bottom coding applied at the level used in the 2010 census are insufficient to prevent re-identification given the ability to perform database reconstruction and the availability of external data.

1.1.2 The use of Differential Privacy

- The proposed use by Census of Differential Privacy to prevent re-identification is promising, but there is as yet no clear picture of how much noise is required to adequately protect census respondents. The appropriate risk assessments have not been performed.
- The Census has not fully identified or prioritized the queries that will be optimized for accuracy under Differential Privacy.
- At some proposed levels of confidentiality protection, and especially for small populations, census block-level data become noisy and lose statistical utility.
- Currently, Differential Privacy implementations do not provide uncertainty estimates for census queries.

1.1.3 Stakeholder response

- Census has not adequately engaged their stakeholder communities regarding the implications of Differential Privacy for confidentiality protection and statistical utility.
- Release of block-level data aggravates the tension between confidentiality protection and data utility.
- Regarding statistical utility, because the use of Differential Privacy is new and state-of-the-art, it is not yet clear to the community of external stakeholders what the overall impact will be.

1.1.4 The pace of introduction of Differential Privacy

- The use of Differential Privacy may bring into conflict two statutory responsibilities of Census, namely reporting of voting district populations and prevention of re-identification.
- The public, and many specialized constituencies, expect from government a measured pace of change, allowing them to adjust to change without excessive dislocation.

1.2 Recommendations

1.2.1 The re-identification vulnerability

- Use substantially equivalent methodologies as employed on the 2010 census data coupled with probabilistic record linkage to assess re-identification risk as a function of the privacy-loss parameter.
- Evaluate the trade-offs between re-identification risk and data utility arising from publishing fewer tables (e.g. none at the block-level) but at larger values of the privacy-loss parameter.

1.2.2 Communication with external stakeholders

- Develop and circulate a list of frequently asked questions for the various stakeholder communities.
- Organize a set of workshops wherein users of census data can work with differentially private 2010 census data at various levels of confidentiality protection. Ensure all user communities are represented.
- Develop a set of 2010 tabulations and microdata at differing values of the privacy-loss parameter and make those available to stakeholders so that they can perform relevant queries to assess utility and also provide input into the query optimization process.
- Develop effective communication for groups of stakeholders regarding the impact of Differential Privacy on their uses for census data.
- Develop and provide to users error estimates for queries on data filtered through Differential Privacy.

1.2.3 Deployment of Differential Privacy for the 2020 census and beyond

- In addition to the use of Differential Privacy, at whatever level of confidentiality protection is ultimately chosen, apply swapping as performed for the 2010 census so that no unexpected weakness of Differential Privacy as applied can result in a 2020 census with less protection than 2010.
- Defer the choice of the privacy-loss parameter and allocation of the detailed privacy budget for the 2020 census until the re-identification risk is assessed and the impact on external users is understood.
- Develop an approach, using real or virtual data enclaves, to facilitate access by trusted users of census data with a larger privacy-loss budget than those released publicly.

- Forgo any public release of block-level data and reallocate that part of the privacy-loss budget to higher geographic levels.
- Amid increasing demands for more granular data and in the face of conflicting statutory requirements, seek clarity on legal obligations for protection of data.

2 INTRODUCTION

2.1 Overview of the Census

The US decennial census, codified in law through the US Constitution has taken place every 10 years since 1790. The 24th such census will take place in 2020. The authority to collect and analyze the information gathered by the Census Bureau originates in Title 13 of the US Code enacted into law in 1954. Title 13 Section 9 of the US code mandates that neither the Secretary of Commerce or any other employee or officer of the Dept. of Commerce may

> "... use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Dept or bureau or agency thereof to examine the individual reports."

Census employees are sworn to uphold the tenets of Title 13 and there are strict penalties including fines and imprisonment should there be any violation. To ensure the mandate of Title 13 is upheld, the Census has traditionally used what are termed Disclosure Avoidance techniques on its publicly released statistical products. The particular approaches used by the Census for Disclosure Avoidance have evolved over the years. A short overview is contained in this report.

Surveys have long been an invaluable tool in determining policy and in the performance of social science and demographic research. In many cases such surveys require respondents to reveal sensitive information under the promise that such information will remain confidential. Traditionally, protection from disclosure was accomplished by anonymizing records. In this way, statistical analyses on issues of public importance could be accomplished while protecting the identity of the respondent. Over time however, the availability of public external data

and the increase in capability of data analytics has made protecting confidential data a challenge. By linking information in one data set with that of another containing some intersecting information (known as a record-linkage attack) it is sometimes possible to connect an anonymous record containing confidential information with a public record and thus identify the respondent. This is called re-identification of previously de-identified data. A number of newsworthy re-identifications have been accomplished in this way. Several approaches have been put forth to make such record linkage attacks harder (see e.g., [32]) but to date none of these have proven to be sufficiently robust to attack.

In 2016, analysts at the Census realized that, even though the Census publishes for the most part tabular summaries of its surveys, enough information could be gleaned from the results to correctly reconstruct a substantial fraction of the detailed survey responses. By linking this information with commercial marketing databases, the names of the respondents could be ascertained, a putative violation of Title 13.

In response, Census has proposed to utilize methods of formal privacy developed and analyzed in the cryptography community; Census proposes to use the methods of Differential Privacy (DP) [8] to secure the 2020 Census. Census requested a JASON study as part of the process of verifying their assessment of disclosure risk as well as assessing the proposed use of formal privacy approaches. Census' charge to JASON was as follows:

- JASON will examine the scientific validity of the vulnerability that the Census Bureau discovered in the methods it has historically used to protect the confidentiality of respondent data when preparing publications.
- Risk assessment: has the Census Bureau properly assessed the vulnerability?
- Implementing formal privacy requires making explicit choices between the accuracy of publications and their associated privacy loss; users always

want more accuracy, but the Census Bureau must also safeguard the respondents' privacy. How do we represent the trade-offs between privacy loss and accuracy to explicitly represent user choices? Are there other conceptual approaches we should try?

2.2 Overview of the Study

JASON was introduced to the relevant issues through a set of presentations listed in Table 2-1. The briefers were experts both internal and external to the Census Bureau in areas such as disclosure avoidance, demography, and applications of census data such as redistricting. These talks were of high quality and were instrumental in educating JASON on these issues. In addition, members of JASON participating in the study were sworn into Title 13 allowing them to be briefed on information protected under this statute and providing JASON with important insights into the details of 2020 Census and particularly the Disclosure Avoidance system based on DP proposed for 2020. Finally, Census provided with JASON with a rich set of reference materials, some protected under Title 13. Details associated with those materials protected under Title 13 are not included in this report.

2.3 Overview of the Report

In Section 3, we provide a brief overview of the census process, the information that Census is mandated to provide and the associated timeline. We also briefly review the methods that were used for Disclosure Avoidance in the past. In Section 4, we review the work that led Census to conclude that the previous approaches to Disclosure Avoidance were inadequate given the increasing availability of large datasets of personal information. In this context, we discuss the seminal work of Dinur and Nissim [5] leading to what is now called the Fundamental Law of Information Recovery. We also describe some experiments asso-

Speaker	Title	Affiliation
Ron Jarmin	Overview of the Dual Mandate and Legal and Historical Background for Disclosure Avoidance	US Census
Victoria Velkoff	Proposed 2020 Census Data Products	US Census
James Whitehorne	Overview of Redistricting Data Products	US Census
John Abowd	The Vulnerability in the 2010 Census Disclosure Avoidance System (DAS)	US Census
Ashwin Machanavajjhala	Interpreting Differential Privacy	Duke University
Dan Kifer	Design Principles of the TopDown Algorithm	Penn State University
Phil Leclerc	Empirical Analysis of Utility-Privacy Trade-offs for the TopDown Algorithm	US Census
William Sexton	Disclosure Avoidance At-Scale and Other Outstanding Issues	US Census
Cynthia Hollingsworth	How 2020 Census Data Products are Prepared	US Census
Rachel Marks	How 2020 Census Data Products Reflect Data User Feedback	US Census
Ken Hodges	How 2020 Census Products will be used by Demographers	Claritas
Justin Levitt	Uses of 2020 Census Redistricting Data	Loyola University
Tommy Wright	Suitability Assessment of Data Treated by DA Methods for Redistricting	US Census
Kamalika Chaudhuri	Formal Privacy and User-Imposed Constraints	UCSD
Salil Vadhan	Formal Privacy and Data Analysis, Including Invariants	Harvard
Dave van Riper	Differential Privacy and the Decennial Census (via VTC)	U. Minnesota
Danah Boyd	Video Teleconference	Microsoft
Jerry Reiter	Video Teleconference	Duke University

Table 2-1: Briefers for JASON Census study.

ciated with the Dinur-Nissim work that underscore the conclusions of that work. In Section 5, we describe briefly the proposed use of DP as a means of protecting sensitive Census data. DP grew out of the work described above by Dinur and Nissim and then extended by Dwork and her collaborators [7]. DP makes possible statistical queries regarding a dataset to be performed while offering a rigorous bound on the amount one learns about a dataset if one record is deleted, added or replaced. Note that this is not, strictly speaking a guarantee of disclosure avoidance but it does provide in a rigorous way the likelihood of a record linkage attack. It does this by adding specially calibrated noise to the result of a specific query made on the dataset. For queries that involve large populations, the addition of noise does not unduly perturb the statistical accuracy of the query. But as a query focuses on smaller and smaller populations the noise will make it increas-

ingly difficult to infer individual characteristics. An attractive feature of DP is that the level of protection is tunable via the setting of a privacy loss parameter. The value set for the privacy loss parameter is meant to be a policy decision.

In Section 6, we discuss the results of some of the early work performed by Census on applying DP to census data. Census proposes to use DP to process the sensitive microdata and create the standard tabular summaries. Noise will then be added to these summaries to make them differentially private. The assessment of the privacy loss budget to be used has not yet been performed. Census will then use the same reconstruction algorithms it applied on the 2010 census data on the noised tables. This will create synthetic microdata that, in principle, should be safe to publish openly. We discuss some early applications of this approach and the nature of the synthetic data it produces. The proposed use of DP will lead to tension between protecting privacy while providing accurate demographic data for activities like redistricting. In Section 7 we propose some approaches for managing this trade-off. Finally in Section 8 we summarize our findings and recommendations.

3 CENSUS PROCESS

In this section we provide a brief overview of the main products that the Census provides as well as the geographic hierarchy that Census has established to collect the relevant respondent data. We also cover the approach the Census has used to process and summarize the required data. Finally, we discuss the evolving need for preservation of the confidentiality of Census data.

3.1 Census Geographical Hierarchy

The Census organizes the US population via a geographic hierarchy shown in Figure 3-1. At the top of this hierarchy are the national boundaries of the United States and Puerto Rico. Within each state, Census further subdivides the population according to county of residence. Counties are then further divided into tracts, block groups, and finally the lowest gradation of Census geography, the Census block. Census also surveys the households in each block and counts for example the number of residents, whether the resident owns or rents etc. Census also collects data for what are known as Group Quarters. Examples of these are dormitories, prisons, etc. The designations in Figure 3-1 of nation, region, state, county, tracts, block groups, and finally census blocks is called the “central spine" of the census geographic hierarchy. Off this spine are also indicated other important state and local divisions. For these, Census provides geographies that can then be used to determine counts in these regions off the spine. These Census geographies inform the placement of Census blocks so that the counts in these areas can be performed from Census block data.

The distribution of population and the number of households in a census tract, block group or block varies greatly across the nation. A map of the population density from 2010 census data is shown in Figure 3-2. As can be seen, the population density varies from thousands of people per square mile as for example in areas like New York City or Los Angeles down to less that ten people per

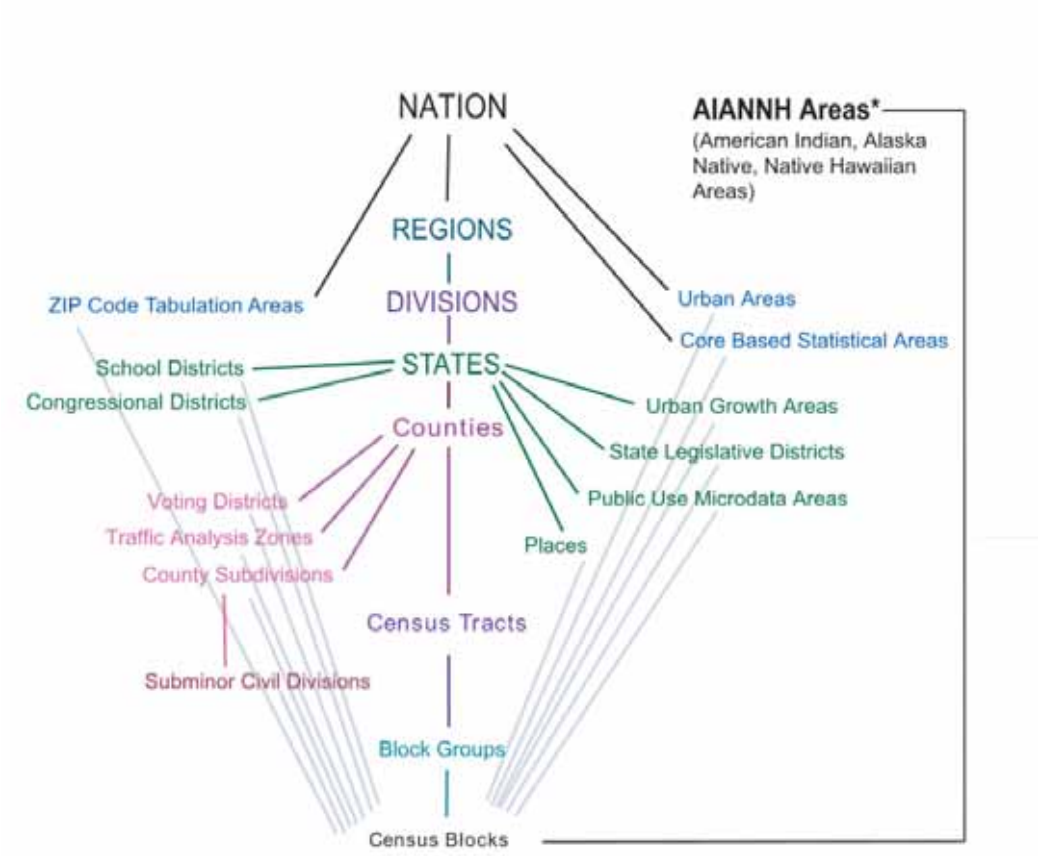


Figure 3-1: The geographical hierarchy used by Census in organizing its various surveys [38].

square mile in states such as Nevada. This diversity in the number of residents and number of households in various regions is one of the reasons Census must work to protect respondent information. In many cases, because of the uniqueness of a given area, it may be possible to identify census respondents. For example, in Figure 3-3 we display graphical representations of the distribution of population and number of households for the country in the form of Violin plots. As can be seen, there is wide variability in both population and number of households even at the census block level. Census blocks are comprised for the most part of roughly several hundred people, but in densely populated areas there are outliers with several thousand people; there is a similar picture for the number of households in a block. Block groups are larger consisting of typically a few thousand

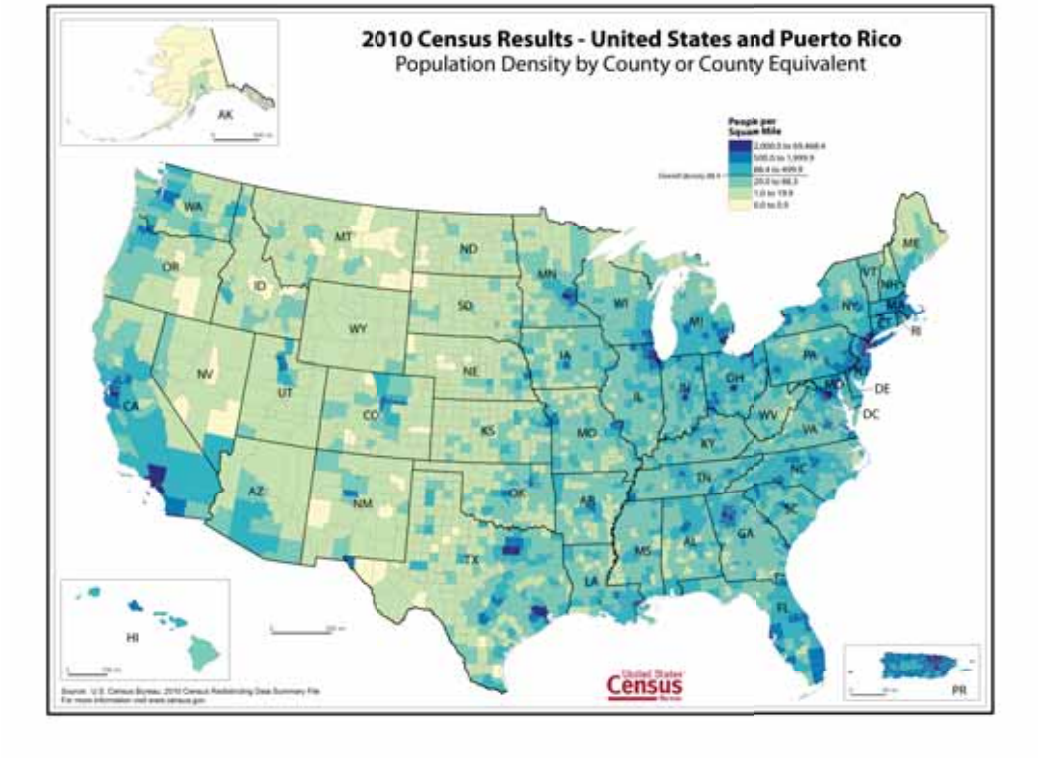
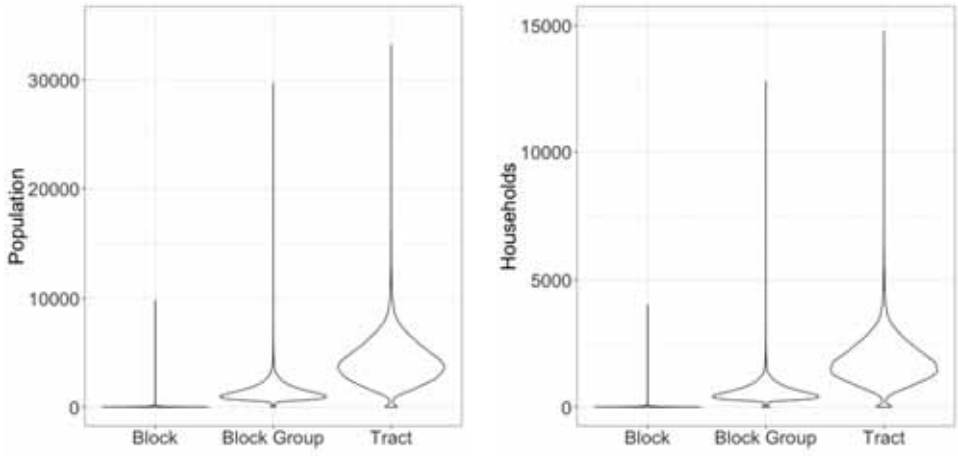
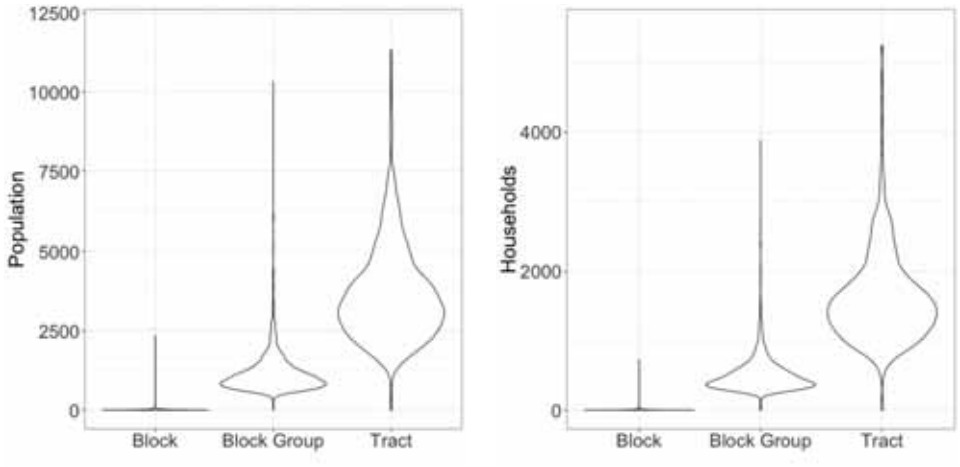


Figure 3-2: Map of population density across the United States from the 2010 census [35].



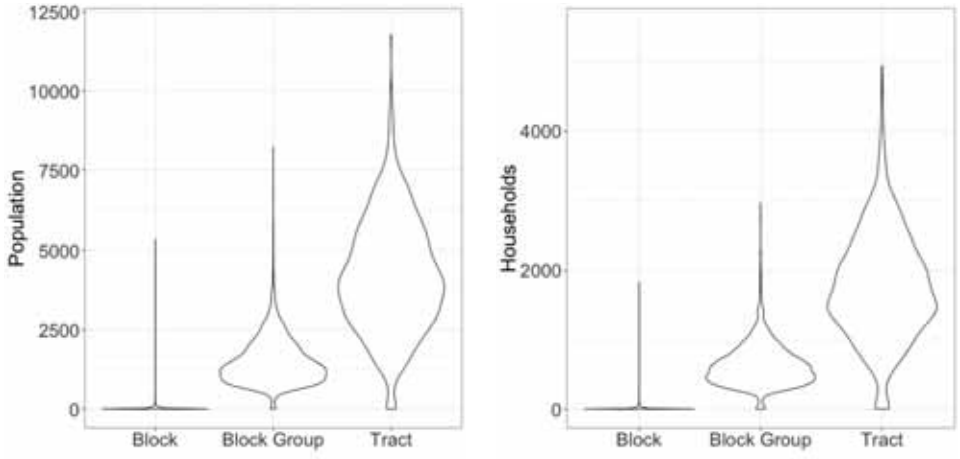
(a) (b)

Figure 3-3: Violin plots of population and households for census tracts, block groups and blocks across the nation.



(a) (b)

Figure 3-4: Violin plots of population and households for census tracts, block groups and blocks in Iowa.



(a) (b)

Figure 3-5: Violin plots of population and households for census tracts, block groups and blocks in Virginia.

people, but here also there is considerable variability. Census tracts may range from population sizes of several hundred in very sparsely populated areas to upwards of 30,000 people. The distribution of population and number of households for blocks, block groups and tracts in a state like Iowa is shown in Figure 3-4. This should be contrasted with the distribution for Virginia shown in Figure 3-5.

Finally it is important to note that census blocks do not always line up with other regions of interest. An important example is the use of census data to determine boundaries of both Congressional and State Legislative districts. Shown in Figure 3-6 are the boundaries for two Congressional districts in Virginia. The boundaries for the districts are shown in black. Census tracts are indicated in purple; census block groups are indicated in orange; and census blocks are indicated in gray. The boundaries for tracts, groups and blocks are quite complex indicative of geography but also complex population patterns. The boundaries of a Congressional district (as well as a state legislative district) are determined through a redistricting process that makes use of the information provided in the PL94 census product (discussed below).

3.2 Census Process and Products

By April 1, 2020 (Census Day) every home will receive a request to participate in the 2020 census. This is the reference data for which respondents report where they usually live. Census then also canvasses group quarters (dorms, etc.) in April. Respondents indicate

- The number of people who live and sleep in a residence most of the time; the homeless are asked to respond as well,
- The ownership status of the household,
- Sex of the residents of the household,
- Age of the residents and their date of birth,
- Whether the residents are of Hispanic origin, [1]
- Race of the residents. This can be any or all of the 63 possible races as designated by the Office of Management and Budget (OMB).

[1] Census refers to this information as the Hispanicity of the respondent.

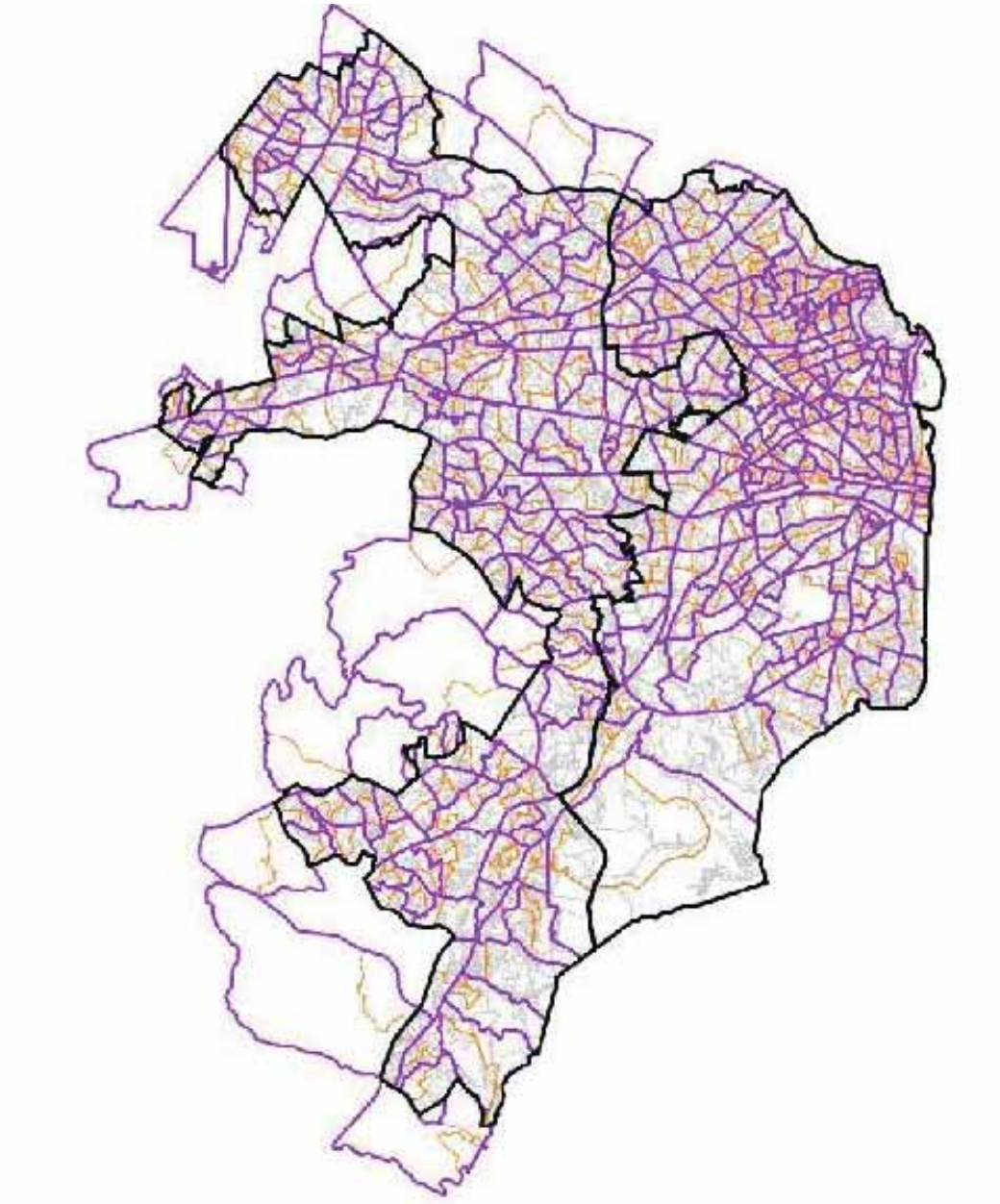


Figure 3-6: A map of two adjoining Congressional districts in Virginia. The black lines indicate the district boundaries; the purple lines indicate boundaries of census tracts; the orange lines indicate boundaries of block groups; the gray lines indicate census blocks.

The 2020 census will also collect information about US citizenship, but respondents will not be asked to indicate their citizenship on the census questionnaire. Instead this will be inferred from existing administrative records (e.g. Social Security Administration, Internal Revenue Service, etc.).

The respondent data are collected into a set of what Census terms microdata, a list of records indicating the responses for each resident. As the responses are received, records are de-duplicated and addresses are validated to insure that every person is counted only once. This forms the Census Unedited File or CUF. Where data are missing or inconsistent the Census employs a process known as imputation and edits the CUF to produce the hundred percent detail file or HDF. The final step is to identify those cells in the various tabular summaries where it may be possible to identify respondents. Here the Census performs confidentiality edits and swaps households as discussed further in Section 3.3. From here the various tabular summaries would be produced.

The Census Bureau through its surveys is responsible for the following products:

Apportionment count Apportionment is the process of dividing the 435 seats of the House of Representative among the states. The count is based on the resident population (both citizen and non-citizen) of the 50 states. An example of the result from the 2010 Census is shown in Figure 3-7 and must be delivered to the President and Congress by December, 2020.

PL94-171 Public law 94-171 directs the Census Bureau to provide redistricting data for the 50 states. This is the first product that must be produced after the apportionment count is complete. Within a year of the 2020 census, the Bureau must send data agreed-upon with the states to redraw state congressional and legislative districts. To meet this requirement the Census has set up a voluntary program that makes it possible for states to receive population estimates as well as racial and Hispanicity distributions for areas relevant to the state congressional and legislative election process. An example of the tables provided in this product is shown

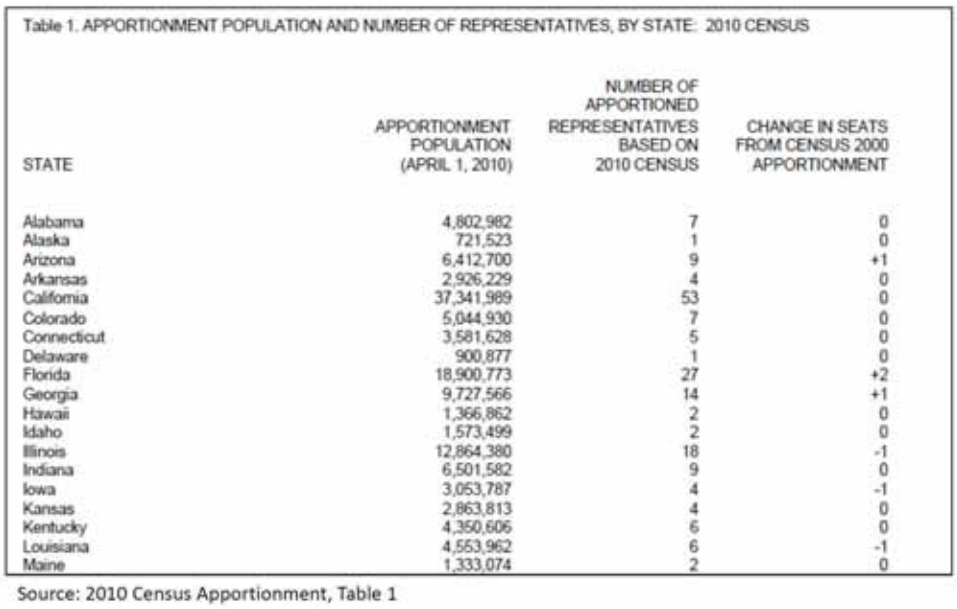


Table 1. APPORTIONMENT POPULATION AND NUMBER OF REPRESENTATIVES, BY STATE: 2010 CENSUS

STATE	APPORTIONMENT POPULATION (APRIL 1, 2010)	NUMBER OF APPORTIONED REPRESENTATIVES BASED ON 2010 CENSUS	CHANGE IN SEATS FROM CENSUS 2000 APPORTIONMENT
Alabama	4,802,982	7	0
Alaska	721,523	1	0
Arizona	6,412,700	9	+1
Arkansas	2,926,229	4	0
California	37,341,989	53	0
Colorado	5,044,930	7	0
Connecticut	3,581,628	5	0
Delaware	900,877	1	0
Florida	18,900,773	27	+2
Georgia	9,727,566	14	+1
Hawaii	1,366,862	2	0
Idaho	1,573,499	2	0
Illinois	12,864,380	18	-1
Indiana	6,501,582	9	0
Iowa	3,053,787	4	-1
Kansas	2,863,813	4	0
Kentucky	4,350,606	6	0
Louisiana	4,553,962	6	-1
Maine	1,333,074	2	0

Source: 2010 Census Apportionment, Table 1

Figure 3-7: A partial list of the apportionment count determining the number of Congressional representatives from each state [39].

in Figure 3-8.

Summary File 1 Census produces a set of demographic profiles after the apportionment and redistricting reports are complete. Summary File 1 (SF1) provides population counts for the 63 OMB race categories and Hispanicity down to the census block level. The report contains data from questions asked of all people and about every housing unit and includes sex, age, race etc. The report consists of 177 population tables, 58 housing tables down to the block level as well as tabulations at the county and tract level. SF1 also provides special tabulations for areas such as metropolitan regions, Congressional districts, school districts etc.

Summary File 2 Summary File 2 (SF2) contains cross-tabulations of information on age, sex, household type, relationship, size for various races as well as Hispanicity down to census tract level as long as the population in the tract exceeds 100 people.

	Virginia	Block 1000, Block Group 1, Census Tract 2001.02, Alexandria city, Virginia	Block 1001, Block Group 1, Census Tract 2001.02, Alexandria city, Virginia
Total:	8,001,024	0	658
Population of one race:	7,767,624	0	630
White alone	5,486,852	0	225
Black or African American alone	1,551,399	0	176
American Indian and Alaska Native alone	29,225	0	3
Asian alone	439,890	0	132
Native Hawaiian and Other Pacific Islander alone	5,980	0	0
Some Other Race alone	254,278	0	94
Two or More Races:	233,400	0	28
Population of two races:	214,276	0	27
White; Black or African American	62,204	0	1
White; American Indian and Alaska Native	25,771	0	0
White; Asian	59,051	0	3
White; Native Hawaiian and Other Pacific Islander	2,618	0	1

Figure 3-8: An example of a population table in the PL94-171 summary file [39].

American Community Survey The American Community Survey (ACS) is an ongoing survey that has taken the place of the decennial long form. It is performed annually. Each year Census contacts 3.5 million households and asks that they fill out a detailed questionnaire. The survey is far more extensive than the decennial census and gathers information about household makeup, type of housing, citi-

zenship, employment etc. The information is used by a variety of stakeholders. Perhaps most importantly, the data are used to guide the disbursement of federal and state funds.

Public Use Microdata Sample Census provides edited samples of the microdata records that make up the decennial census and the ACS. These records are assembled for areas that contain a minimum population of 100,000 (known as PUMAs) and are edited to protect confidentiality. The PUMS provides only a 10% sample of a PUMA.

3.3 The Need for Disclosure Avoidance

It was realized early on that some disclosure avoidance was necessary as the population and housing densities of the United States are not distributed in a homogeneous manner. Owing to special aspects of a location it may be possible to identify the particular person or persons living there. This would constitute a violation of Title 13. For example, Liberty Island, the base of the Statue of Liberty has one household listed, that of the Superintendent of the Monument and his wife [13]. Thus by focusing on this location and using external sources it should be possible to identify the residents of that particular household. For this reason, the information for this location is swapped with that of another household. A history of the methods used in the past 50 years to effect disclosure avoidance is available in the paper by McKenna [24]. We briefly describe these here to provide some context for this report. The discussion below is not complete but illustrates the evolution of the need to offer improved disclosure avoidance.

Long form data Long form census data have never been published at the lowest level of census geography (presently census blocks). The long form data were generally collected as part of the decennial census but in 2010 this data was relegated to what is now called the American Community Survey (ACS) which began

in 2005. The ACS only publishes data down to the block group level.

1970 Census The 1970 Census utilized suppression of whole tables as opposed to suppression of cells. The choice to suppress was based on the number of people in households in a given area. This approach had limitations in that tables with complementary information were not suppressed making it possible in some cases to infer the suppressed information. As indicated by McKenna, cells within an original table could still show an estimate of 1 or 2 people.

1980 Census The 1980 Census retained the approach of the 1970 census but modified it further by now suppressing tables with complementary information and zeroing cells with counts of 1 or 2. However some population counts were not suppressed at any level. In some cases, one could still infer complementary data by subtracting data for various counties from state populations to infer population results for a county that had been suppressed.

1990 Census The 1990 census was the first to employ the concept of swapping. The 100% data (namely PL94, Summary File 1 and Summary File 2) were published down to the block level. But, where there was risk of potential disclosure, a confidentiality edit was performed on the census microdata. For those small blocks deemed at risk, Census selected a small sample of households with a higher sampling rate of such at-risk households used in small census blocks. These at-risk records were paired with other census records from other geographic locations using a set of matching rules. The matching process preserved key attributes such as household size, the age of those residing in a given location, etc. The household records are then swapped and the interchanged version is what is used for the Census Edited File that then forms the source of the various tabular summaries. The rate of swapping is not disclosed so as to prevent possible reverse engineering of the process. In addition, Census began using rounding of entries as well as top and bottom coding to prevent respondent identification arising from

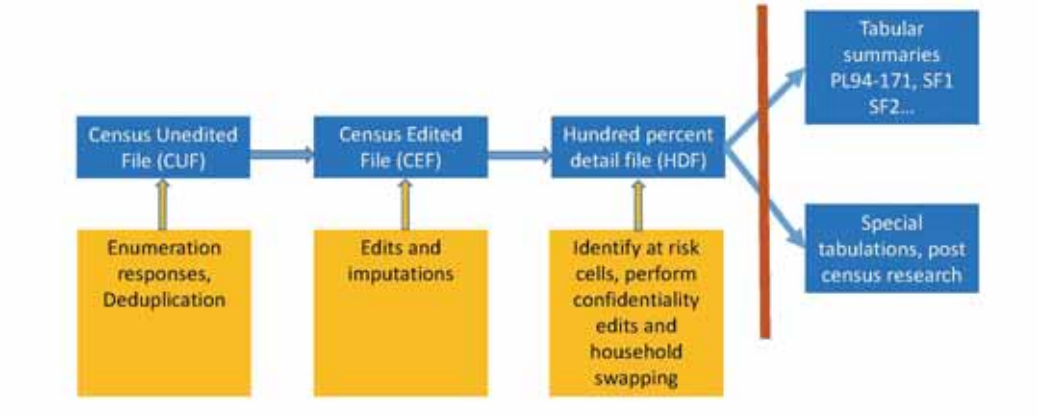


Figure 3-9: A graphical depiction of the disclosure avoidance process used in the recent 2010 census.

age extremes etc.

2000 Census For the 2000 census, more emphasis was given to protecting small blocks and block groups from possible re-identification. For this census, the race category was expanded to include 63 possible alone or combined races. The probability of swapping was increased to those cases where disclosure risk was thought to be higher such as cross-tabulations of key variables, smaller blocks, and also households that contained unique races in that census block.

2010 Census The approach to disclosure avoidance used in 2010 largely followed the approaches developed in the earlier 1990 Census as discussed above. In addition, Census developed partially synthetic data for group quarters in which it blanked values that were assessed as at risk and instead substitutes those values with data obtained from regression models. In summary the disclosure avoidance process follows steps outlined graphically in Figure 3-9. In the next section we discuss why this approach was ultimately judged inadequate.

4 THE CENSUS RE-IDENTIFICATION VULNERABILITY

In this section we discuss the vulnerability discovered by Census using the 2010 census data. We then examine the fundamental basis of the vulnerability: the results demonstrated in 2003 by Dinur and Nissim [5] that releasing an overly large number of statistics about a database allows one to perform reconstruction of the detailed data comprising that database. This result holds true even when one tries to preserve privacy by noising the results of database queries. We verify some of their observations in this section. We also offer a reinterpretation of their results in terms of information theory. Our discussion essentially validates the conclusion of Census that it is possible to reconstruct census microdata even after the application of traditional disclosure avoidance techniques like swapping, top and bottom coding etc.

4.1 Reconstruction of Census Tabular Data

The tabular summaries found in Census products such as PL94-171, SF1 and SF2 have been viewed in the past as safe to publish. These summaries are built using census microdata and it is this microdata that is controlled via disclosure avoidance. For the 2010 census the techniques discussed in Section 3.3 were all used; randomized swapping of households, top and bottom limitations on populations and ages, etc.

In 2018 Census looked at the feasibility that the tabular summaries could be processed to infer the microdata records that were used to produce them [1]. This had not been thought to be feasible owing to the large amount of data and computation involved. Such reconstruction of the microdata is not yet a violation of Title 13 since no personal data (e.g. names, addresses, etc.) are used when these tables are built. But, as in other re-identification attacks, if external data can be joined with the microdata then it may be possible to relink the microdata with

the associated personal data.

In creating the major products published by the Census, each time a cell is populated in a table it is a result of a query made on the microdata. For 2010 the number of queries (or equivalently the number of tabulations) in the PL94 publication is about 3.6B or about 10 for every person in the US. For SF1, the number of tabulations is 22B for population and 4.5B for tabulations of households or group quarters. For SF2 there are 50B tabulations. And for the survey of American Indians and Alaskan Natives there are 75B tabulations. Thus Census publishes a total of 155B queries over the population and households of the US. The population of the US in 2010 was approximately 310M and so many more queries than people (by a significant multiple) have been issued. Most of the microdata entries used to produce these tables have not been processed through traditional disclosure methods.

To test the likelihood of reconstruction Census selected only a subset of the tables that are published. These were

P001	Total population by block,
P006	Total races tallied by block,
P007	Hispanic or Latino origin by race by block,
P009	Hispanic or Latino and not Hispanic or Latino by race by block,
P011	Hispanic or Latino and not Hispanic or Latino by race by age (≥ 18) by block,
P012	Sex by age by block,
P012A-I	Sex by age by block iterated by race,
P014	Sex by age (< 20) by block,
PCT012012A-N	Sex by age by tract iterated by major race alone.

Each table entry is equivalent to an integer-valued linear equation over the microdata tables. For example, if we set the count of people in tract t who are

male and who are 27 years old to $T_{t,M,27}$ then this is tabulated via the equation

$$T_{t,M,27} = \sum_p \sum_r \sum_b B_{p,M,27,r,b}, \tag{4-1}$$

where p sums over the internal person number in the microdata, r sums over the possible races, and b sums over the block codes associated with tract t. The summand B is a selector that is 1 if a record indicates a male of age 27 of any race residing in a block in tract t and zero otherwise [17]. The sum over race is necessary to pick up one of the 63 combinations of race recognized by OMB.

To solve the resulting collection of equations, Census used a state of the art optimization solver known as Gurobi [12]. The Gurobi solver attempts to find the best integer solution to the set of equations corresponding to the tabulations. To break up the problem into manageable pieces Census applied the solver at the tract level. The solver was able to solve the resulting systems with few exceptions. The microdata for the entire US was determined in this way for all 70,000 Census tracts and all 11M Census blocks. To perform the relevant calculations, a virtual parallel cluster was instantiated using Amazon Elastic Cloud facilities and, for this workload and cluster configuration, completed the task in several weeks. Such a task therefore is not outside present day capabilities.

The resulting reconstructed microdata contained

- A geocode at the block level
- A binary variable indicating Hispanic origin (or not) and one of the 63 possible OMB race categories
- Sex
- Age (by year).

Census does publish a sample of the microdata called the Public Use Microdata Sample (PUMS) for use by demographers and other researchers for both the decennial census and for the American Community Survey, but these are rigorously

curated to make sure individual information cannot be inferred. For example, the geographic resolution is limited to areas with populations over 100000. In contrast, the reconstructed data has no population threshold and contains data like single year ages, race, and ethnicity at the block level.

The next step was to see if the reconstructed microdata could then be linked with commercially available marketing data. Some of this data is freely available or could be reconstructed using public records, but more complete and current databases can be licensed through marketing research firms. Such commercial data typically contain names, addresses, sex and birthdate but typically do not contain information regarding race and ethnicity. While not investigated in this case, Census data also contain information about family make-up. Using the reconstructed database, and acquiring commercial data, Census performed a database join using the age, sex and block locations as the common columns of the two datasets. The entries in the resulting table would now have the name and address of the respondent. If correct, these would be a re-identification of the microdata records. Release of this information would constitute a violation of Title 13.

Census determined that 46% of the reconstructed records matched correctly to the internal microdata. If a fuzzy match on age were used, 71% of the records matched. Thus the reconstruction algorithm using only some of the Census tables matched correctly 71% of the US population. Of those internal Census records, 45% were successfully mapped to a corresponding record in a commercial database again using fuzzy age matching with a one year uncertainty. Census then took the records that matched to see if they in turn matched the internal records Census collects when people submit their responses that contain name and address. Of the records that matched the commercial data sets, 39% of these matched exactly with Census records. This corresponds to the successful re-identification of 52M people or 17% of the population in 2010. Previous estimates of the re-identification rate was 0.017% of the population and only 22% of these were confirmed to be correct. The re-identification risk demonstrated by Census is four orders of magnitude larger than had been previously assessed [27].

In section 4.2 we examine a simplified version of this reconstruction problem in which the data set is just a column of bits to verify that the type of attack described above is not specific to the data protected by the Census. It is a general difficulty associated with publishing too many query results about a sensitive dataset.

4.2 Results of Dinur and Nissim

As discussed in Section 4, a key motivation for the development of formal privacy approaches to further secure the 2020 census is the Fundamental Law of Information Recovery. This observation, as quoted by Dwork is that

> "overly accurate estimates of 'too many' statistics is blatantly nonprivate."

By blatantly nonprivate is meant that given some database with information we wish to keep private there exists a methodology to issue queries on the dataset that will allow one to infer a dataset whose elements differ from the original in some number of elements. The number of elements that are not obtained correctly reduces as the size of the database increases. Thus for a large enough database the methodology asymptotically extracts all the elements of the private database.

Dinur and Nissim [5] demonstrated this in a seminal paper by modeling a database as a set of binary numbers whose (private) values we are interested in learning. The database is represented by an array of binary digits:

$$d = (d_1, d_2, \ldots, d_n). \tag{4-2}$$

A *statistical query* is represented by a subset $q \in [1, 2, \ldots, n]$. The exact answer to the query is the sum of all the database entries specified by q:

$$a_q = \sum_{i \in q} d_i. \tag{4-3}$$

An answer $A(q)$ is said to be within ε perturbation if

$$|a_q - A(q)| \leq \varepsilon$$

The *algorithm A* is said to be within ε perturbation if for all the queries $q \subseteq [n]$ the answers A are within ε perturbation. Dinur and Nissim define the notion of $T(n)$ non-privacy if there exists a Turing machine that terminates in $T(n)$ steps so that the probability of determining any fraction of the bits with the exception of a vanishingly small number as the size of the data set increases is essentially one. The result of most relevance to this study is that if the query algorithm provides $o(\sqrt{n})$ perturbation then non-privacy can be achieved with an algorithm that terminates in a number of steps that grows polynomially with increasing data set size. More noise than this is required to get even weak privacy. Dinur and Nissim describe an algorithm using linear programming to demonstrate the existence of such an algorithm. The conclusion is that, even in the presence of noise, a sufficiently capable adversary can infer the secret bits of the dataset. In order to ensure privacy one must restrict the number of queries or add so much noise that the utility of statistical queries on the dataset is potentially degraded.

4.3 JASON Verification of the Dinur-Nissim Results

JASON undertook a verification of the Dinur-Nissim results using a variation of their approach. First we examine the situation where no noise is added to the queries. We then examine the situation where we add noise. We begin by generating a random vector of zeros and ones, $\boldsymbol{d}$, of size n. We then create an $m \times n$ random matrix, Q of zeros and ones. These will be the queries. We then compute the matrix vector product of the query matrix with the database vector. These are the random query results. We then use bounded least squares with constraints to solve the following problem:

$$\text{argmin} \, ||Q\boldsymbol{x} - \boldsymbol{d}||^2 \text{ subject to } 0 \leq x_i \leq 1. \tag{4-4}$$

Once this problem is solved we then round the components of the resulting vector

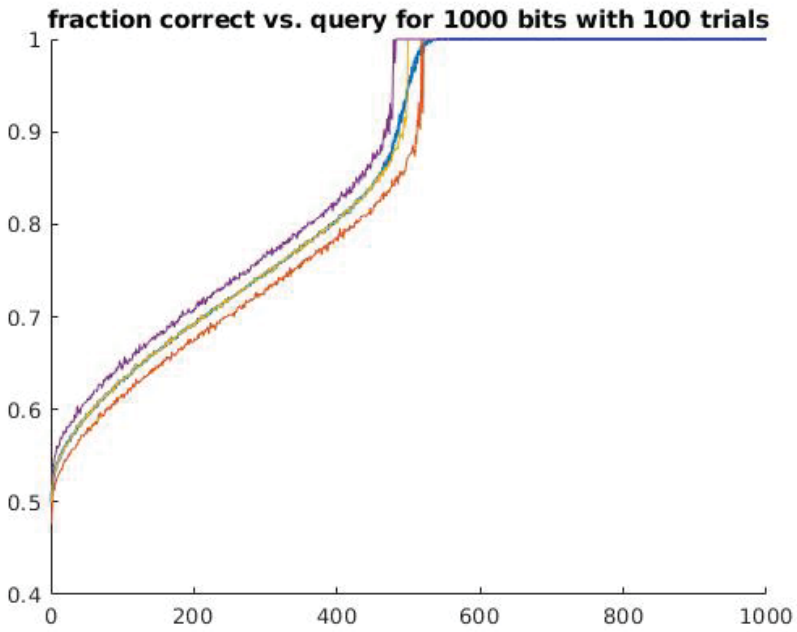


Figure 4-1: Fraction of bits recovered for a 1000 bit Dinur-Nissim dataset as a function of the number of random queries. The lower curve is the minimum fraction recovered, the middle curve is the mean, and the upper curve is the maximum recovered. No noise is added to the query results.

x to 0 or 1. If we issue n queries and our query matrix is not singular,[2] then we would recover the results of the database immediately. But in fact the full database can be recovered with less than n queries in the absence of noise. In Figure 4-2 we plot the fraction of bits computed correctly as a function of the number of queries for a database of size 1000 bits. Because our queries are random we perform 100 trials and plot the 10% decile of the fraction of bits recovered (lower curve), the 90% decile fraction of bits recovered (upper curve) and the mean recovered (middle curve).

With no queries we recover 50% of the bits, but this is of course no better than random guessing. As the number of queries increases we recover more of the bits (although the bits recovered will differ with each random attempt). It is to be expected that we would recover all the bits once we issue 1000 random

[2] singularity would be a very rare event

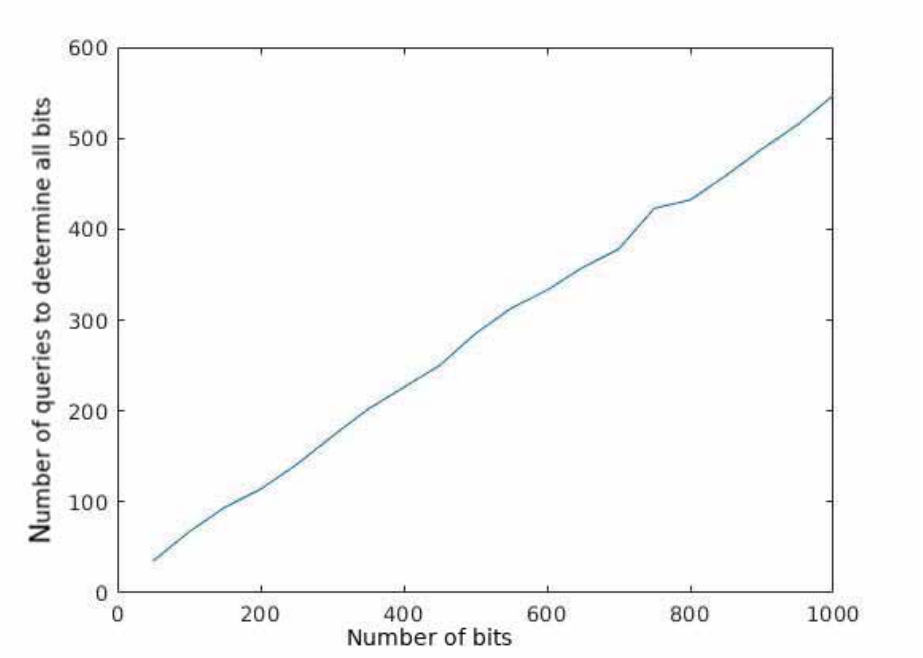


Figure 4-2: Number of queries needed to recover 100% of the private bits in the Dinur-Nissim dataset as a function of the size in bits of the data set.

queries but as is seen in the Figure all the bits are recovered at about the half way mark in the number of queries. If one repeats this calculation for databases of varying size n and asks how the number of queries required to achieve perfect knowledge of the bits varies with n one gets a roughly linear variation in n as shown in Figure 4-2. The slope of this roughly linear variation as a function of increasing database size is shown in Figure 4-3. As can be seen the slope is close to 1/2 indicating that roughly $n/2$ queries are required on average to determine the entire database. This is a special aspect of this particular type of database. A random query response will get information about a number of the bits. For example, if we choose to query two bits at a time by summing the values, then a sum of zero immediately tells us the two bits must be zero. Similarly if we get a sum of 2 we know immediately the two bits we queried must have both been one. Thus one can infer the bits more quickly in a probabilistic sense then simply asking for one bit at a time which would correspond to the query matrix being the identity. In section 4.5 we apply an information-theoretic argument to show that

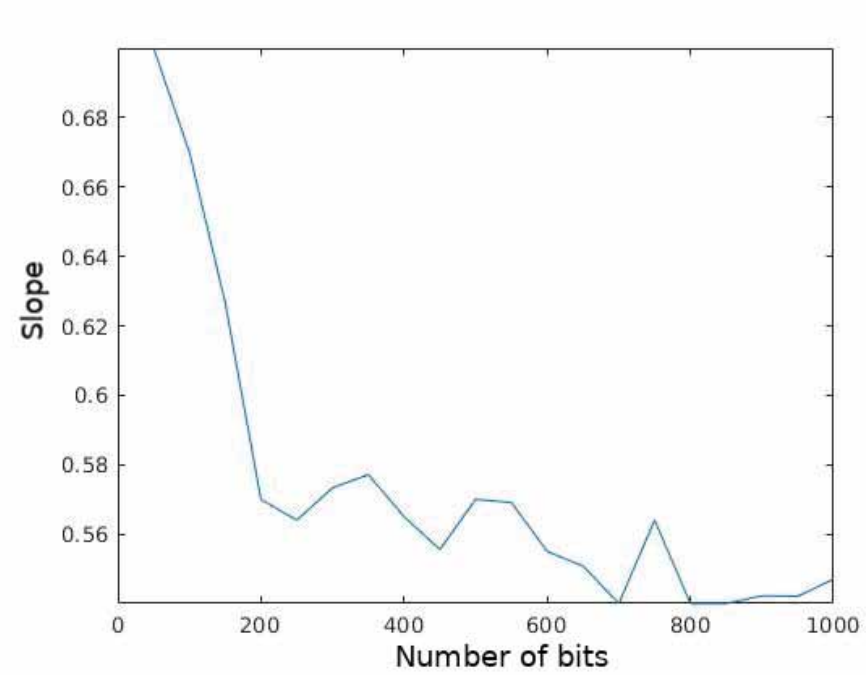


Figure 4-3: Same as Figure 4-2 but each point is normalized by the number of queries. As the number of of bits increases the curve appears to approach a limit of 1/2

the results we get from our least squares approach are not far from optimal.

The results above certainly confirm that, without noise, it is possible through a sequence of queries to infer the entries of a database. It should also be noted that a recovery approach based on optimization will also succeed if one poses more queries than the number of entries in the database. To be sure, the Dinur-Nissim database is special, but it is easily confirmed that through publication of tabular summaries that comprise (sometimes multiple times) the information contained in the database, recovery of the bits, in this case a stand-in for microdata, is possible.

If we think of census data as a (very large) Dinur-Nissim database we can see that the reconstruction attack is quite plausible. In terms of bits, a rough count of the number of bits contained in the Census Edited File might be

- 3 bits to describe the 8 types of group quarters (8 levels),

- 5 bits to describe a person's age (here we assume ages are only reported in intervals of 5)
- 1 bit to describe Hispanic origin (2 levels),
- 6 bits to describe race (63 OMB race designations),
- 24 bits to describe the 11 million census blocks,

for a total of 39 bits per person. If we estimate that in 2010 there were 3×10^8 residents in the US this totals to 1.2×10^{10} bits. If we examine the number of queries in a full cross table this would be

$$(8 \times 20 \times 2 \times 63) \times 1.1 \times 10^7 = 2.2 \times 10^{11}$$

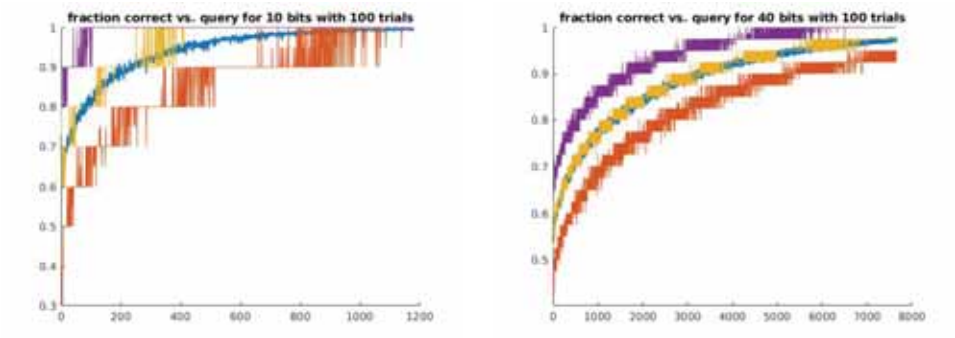
This rough estimate indicates that the census tables "overquery" the data set by a factor of almost 20. If we treat the Census database reconstruction effort as an attempt to infer the bits in a large Dinur-Nissim database there is no question the database (up to the edits that are used to create the tables) could be reproduced with perfect accuracy. A similar argument using the idea of Boolean satisfiability (SAT) solvers is given in [10].

4.4 Queries in the Presence of Noise

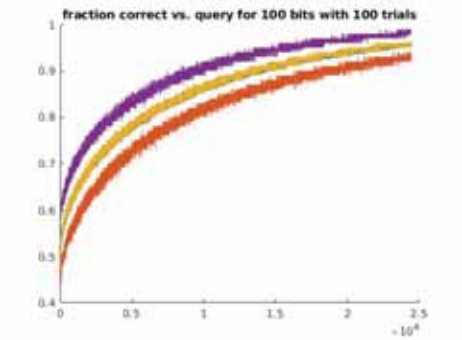
Given the vulnerability discussed above it is perhaps of more interest to examine the number of queries that must be issued to recover the database when each query is perturbed by noise. To examine this, we used the same bounded least squares optimization approach but in the presence of noise. For a dataset size of n bits we added to each random sum a perturbation sampled from a normal distribution of mean 0 and variance $\sqrt{n}\log(n)/2$ where n is again the number of secret bits in the database. The reason for this particular choice was to see if the optimization approach would fail with an increasing number of queries. According to Dinur and Nissim if one adds noise with an amplitude of greater than $O(\sqrt{n})$ then recovery

should be impossible. We were unable to confirm this observation. Instead, as the number of queries increases, an increasing fraction of the correct bits is returned. This is most likely not in conflict with the theorems of Dinur and Nissim as they require that the adversary be time bounded whereas in our approach we do not impose any time limit but instead continually issue queries. The results are shown in Figure 4-4. In the Figure we show the fraction of bits determined correctly as a function of the number of queries for databases of varying size. For each database of size n we added a random perturbation sampled from a normal distribution of mean 0 and variance $\sqrt{n}\log n/2$ to each query.

We perform a query of size m 100 times and provide some statistics for the results. The red, yellow and purple lines indicate the 10%, 50% and 90% deciles respectively of fraction of bits recovered correctly; the blue lines indicate the mean of the fraction of bits recovered correctly. As can be seen, the number of queries required increases greatly, but, in all cases, all metrics measuring the fraction of bits recovered correctly increase towards one. Thus if one is willing to issue a large number of queries, for example, a large multiple of the number of bits, eventually one will learn the internal records of the database. Apparently, the use of random queries will provide results that average out the applied noise and recover the required information. In some ways this is to be expected. For example if we were allowed to issue directly a query for bit i of the n bits in the presence of noise, we would have received a random response, but continual averaging over the responses would have recovered the result regardless of the amount of noise. Indeed we would have predicted that we would have required a number of queries which is some constant factor of the variance. We discuss this further in Section 5 where we consider how many queries are required for a given noise level to recover the internal bits. In the next section we apply information theory to compute idealized estimates of the number of queries required to infer the internal data of the Dinur-Nissim database both in the absence and presence of noise.



(a) (b)



(c)

Figure 4-4: Fraction of bits recovered as a number of queries for databases of size 10, 40, and 100 bits. For each case we infuse the query results with Gaussian noise of means 0 and variance $\sqrt{n}\log n/2$. The red, yellow and purple lines indicate the 10%, 50% and 90% deciles respectively of fraction of bits recovered correctly; the blue lines indicate the mean of fraction of bits recovered correctly. Note that as the number of queries increase, the fraction of bits recovered grows until all the bits are recovered with near certain probability.

4.5 Information Theory and Database Uniqueness

The purpose of this subsection is to look at Dinur & Nissim's [5] fundamental results about database reconstruction from alternative points of view, namely linear algebra and (especially) information theory. The discussion is rather lengthy (but we hope pedagogical) so we have relegated it to an Appendix, but we summarize the main results:

1. In the absence of noise, a database of $n \gg 1$ bits is determined by the re-

sults of approximately $2n/\log_2 n$ queries, on the average over all possible databases. Put differently, we can expect to recover most of the bits of most databases.

2. If noise with variance $\sigma_N^2 < n/48$ is added to the results of each query, the database remains determined by no more than $\sim n$ queries on average.

3. If the noise variance $\sigma_N^2 \gg n/16$, we expect to require $\sim 16\sigma_N^2$ queries to fix the bits uniquely.

It should be noted that there are at least two facets to DN's results: (i) $o(\sqrt{n})$ noise allows the database to be uniquely specified using algebraically (in n) many queries; and (ii) the bits can actually be reconstructed in polynomial time using linear programming. Apart from a few obvious remarks about linear algebra in the noiseless case, we have nothing to say here about the computations required to do the actual reconstruction. Our information-theoretic arguments advanced here are nonconstructive, in much the same way as the Shannon channel-capacity theorem [31], which does not say by what encodings the capacity can be achieved.

5 DIFFERENTIAL PRIVACY

The Census has proposed the use of Differential Privacy (DP) as the basis for its future Disclosure Avoidance System (DAS). The goal of DP is to prevent one from learning about the possible participation of an individual in a survey. The idea is that the result of a query into the dataset provides results that are largely the same even if an individual opted out of participating in the survey. This is accomplished by adding noise to the results of queries so that one cannot easily perform the types of record linkage attacks that have determined the details of database records from queries in the past. DP introduced by Cynthia Dwork [7, 8] and colleagues and developed since then in a vast research literature is viewed as the present gold standard for formal privacy guarantees. The definition is phrased in a language that may be unfamiliar, so we go over it in detail.

The setting is databases and database queries. A database D is a collection of records. Each record has attributes (age, sex, HIV-positive, wealth, or whatever), and each attribute has a range of values it can take. A query is just some function on the database. For instance, "how many records are there", "what is the average age of HIV-positive people", and so forth. We think of attributes being exact and queries giving precise answers, but that is not always desirable as we have discussed previously and is in fact a mental shortcut. Age is reported in years, not days, so people with age 12 are those aged between 12 and 13. Then average age is also reported in years, not some exact number like 62381/129.

DP is a property of algorithms for answering queries. It is clear that, to preserve privacy, queries cannot just return the right answer, so one can think of an algorithm that answers a query as adding noise to the correct answer. Adding noise means that the algorithm is not deterministic, but probabilistic, using random numbers. The approach in which noise is added to the query is known as a mechanism.

An algorithm $\mathcal{A}$ is ε-DP (ε-differentially private) if

$$e^{-\varepsilon} < \frac{\Pr(\mathcal{A}((D)) \in T)}{\Pr(\mathcal{A}(D') \in T)} < e^{\varepsilon}$$

where D and D' are any two databases that differ by one record. The probabilities come from the random numbers that $\mathcal{A}$ uses. T is the set of possible outcomes of $\mathcal{A}$. For instance, if the query was for average age, then T would be an interval like [37,38), meaning that the average age is between 37 and 38. Alternately, if $\mathcal{A}$ returns continuous values, then one needs to measure the probability that the result lies in an interval, rather than takes on a specific value.

A key element of DP is the notion of the privacy budget. In the DP literature this is typically labeled ε. The notation is set up so that a value of $\varepsilon = 0$ indicates zero privacy loss. The technical definition of a DP algorithm is as follows:

Theorem. *An algorithm $\mathcal{A}$ satisfies differential privacy if and only if for any two datasets D and D' that differ in only one record, we have that for all results T that lie in the range of the algorithm $\mathcal{A}$*

$$Pr[\mathcal{A}(D) \in T] \leq \exp(\varepsilon) Pr[\mathcal{A}(D') \in T].$$

Equivalently the ratio of probabilities

$$\frac{Pr[\mathcal{A}(D) \in T]}{Pr(\mathcal{A}(D') \in T} \leq \exp(\varepsilon).$$

Note that there is nothing special about D and D' so we can write the inequality in a symmetric two-sided manner as we did above:

$$\exp(-\varepsilon) \leq \frac{\Pr[\mathcal{A}(D) \in T]}{\Pr[\mathcal{A}(D') \in T]} \leq \exp(\varepsilon).$$

If an algorithm satisfies the definition of being differentially private, the expression above provides a bound on how much additional information one can infer from adding or deleting a record in a database. This will prevent learning about a specific record through the examination of the two datasets for example through database differencing. It also makes record linkage attacks more difficult in that it introduces uncertainty in the query results.

Perhaps of more importance, DP algorithms by definition provide formal bounds on how many queries can be made before the probability of learning something specific about a database increases to an unacceptable level. This is the real role of the privacy budget. A DP algorithm with a large value of ε indicates that the ratio of probabilities of learning a specific result in two datasets with one record differing is large and so implying that the query using the algorithm discriminates strongly between the two datasets. On the other hand, a small value of ε means little additional information regarding the dataset is learned. It is not hard to show that DP has several properties that make it possible to reason about how the privacy budget is affected by queries.

Sequential access to the private data degrades privacy Suppose we have an algorithm $\mathcal{A}_1$ that satisfies DP with privacy loss parameter ε_1 and another algorithm $\mathcal{A}_2$ that has a privacy loss parameter ε_2. If both algorithms are composed then the privacy loss parameter for the composed algorithm is the sum of the individual privacy loss parameters. we have

$$\begin{aligned}\Pr[\mathcal{A}_2(\mathcal{A}1(D),D)=t] &= \sum_{s\in S}\Pr[\mathcal{A}_1(D)=s]\Pr[\mathcal{A}_2(s,D)=t]\\ &\leq \sum_{s\in S}\exp(\varepsilon_1)\Pr[\mathcal{A}_1(D')=s]\exp(\varepsilon_2)\Pr[\mathcal{A}_2(s,D')=t]\\ &\leq \exp(\varepsilon_1+\varepsilon_2)Pr[\mathcal{A}_2(\mathcal{A}_1(D')D')=t].\end{aligned}$$

In general, if one composes this way k times the effective ε becomes

$$\varepsilon = \varepsilon_1+\varepsilon_2+\cdots\varepsilon_k.$$

This implies that one must account for all the operations to be performed on the data in order to ensure a global level of privacy over the whole dataset. It also demonstrates, at least in terms of bounds, the cost of a number of queries on a database in terms of overall privacy and that repeated queries on the data will boost the ratio of probabilities. This provides a useful quantitative aspect to assessing disclosure risk atlhough it is not explicitly a statement about disclosure risk.

The privacy budget behaves gracefully under post-processing If an algorithm $\mathcal{A}_1$ satisfies DP with a privacy budget of ε, then for any other algorithm $\mathcal{A}_2$ which post-processes the data generated by $\mathcal{A}_1$, the composition of $\mathcal{A}_2$ with $\mathcal{A}_1$ satisfies DP with the same privacy budget. To see this, suppose S is the range of the algorithm $\mathcal{A}_1$. Then we have

$$\begin{aligned}\Pr[\mathcal{A}_2(\mathcal{A}_1(D)) = t] &= \sum_{s\in S} \Pr[\mathcal{A}_1(D) = s]\Pr[\mathcal{A}_2(s) = t] \\ &\leq \sum_{s\in S} \exp(\varepsilon) Pr[\mathcal{A}_1(D') = s]\Pr[\mathcal{A}_2(s) = t] \\ &\leq \exp(\varepsilon)\Pr[\mathcal{A}_2(\mathcal{A}_1(D')) = t].\end{aligned}$$

It is important in this argument that only the algorithm $\mathcal{A}_1$ accesses the private data of the database. This composition property is quite powerful. One of its most important applications is that if you transform the database into another database with synthetic data processed through a DP algorithm then additional processing of that data will preserve differential privacy. Thus one can create a dataset from the original dataset and preserve differential privacy for future processing of the synthetic data. This feature is an important component of the disclosure avoidance system currently under consideration by Census.

Parallel composition If one deterministically partitions a database into separate parts then one can control the privacy loss. If $\mathcal{A}_1, \mathcal{A}_2 \ldots, \mathcal{A}_k$ are algorithms that respectively only access the (nonoverlapping) partitions of the database D_1, D_2, $\ldots$ D_k then publishing the results of the queries $\mathcal{A}_1(D_1), \mathcal{A}_2(D_2), \ldots \mathcal{A}_k(D_k)$ will satisfy DP but with an ε given by

$$\varepsilon = \max(\varepsilon_1, \varepsilon_2, \varepsilon_k).$$

Such results show that the production of a histogram where the data is partitioned into categories and then counts are published for each category can still preserve a given privacy budget.

5.1 Mechanisms

The definition of DP does not guarantee that there are any DP algorithms, but of course there are. In general, a *mechanism* is a way of generating DP algorithms from data base queries. We discuss some of these below.

5.1.1 Laplace mechanism

Consider a query whose correct answer is some continuous numeric value. The query has sensitivity Δ if the correct answer on any two neighboring databases D, D' can differ by at most Δ. Then an ε-DP algorithm for this query would add $\text{Lap}(\Delta/\varepsilon)$ noise sampled from a Laplace probability distribution to the correct answer, where Lap is the two-sided Laplace distribution. The probability density for the Laplace distribution with parameter β is

$$\frac{1}{2\beta}\exp\left(-|x|/\beta\right).$$

More usefully, to generate a random Laplace variate from a uniformly distributed p between 0 and 1, one can compute

$$\beta\ \text{sgn}(p-0.5)\ln(1-2|p-0.5|).$$

This density has mean 0 and a variance of $2\beta^2$ and is displayed in Figure 5-1. In applications to DP we use the relation $\beta = 1/\varepsilon$. Thus small values of privacy loss imply large values of β and so very broad distributions with large variances. Note that the use of the Laplace mechanism and the associated Laplace distribution matches exactly with the definitions of DP in terms of the bounds on probabilities. Other distributions can be used, for example, a normal distribution, but in this case there may be small violations of the DP bounds for extreme values. A slightly modified definition of DP is required to handle this case but its use would not affect our conclusions so we won't discuss it further.

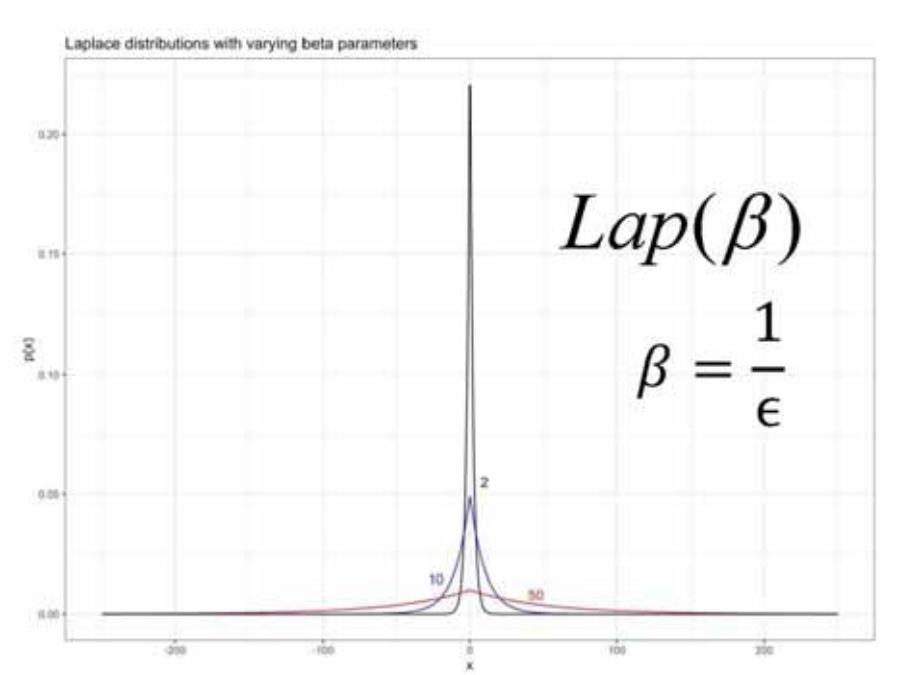


Figure 5-1: The Laplace distribution for several values of the parameter β. A large β corresponds to broad tails.

5.1.2 Geometric mechanism

The Laplace mechanism does not produce integers for integer-valued attributes. The Geometric mechanism adds an integer to the correct answer, where the integer is randomly chosen from a suitable geometric distribution One could instead use the Laplace mechanism and round, but these results are slightly different. The (two-sided) geometric distribution with parameter α has probability density

$$\frac{\alpha-1}{\alpha+1}\alpha^{-|x|}$$

for producing integer x. If Δ is the sensitivity of the query, ε-DP is the same as $\alpha = \exp(\varepsilon/\Delta)$.

5.1.3 Matrix mechanism

In applying DP to the census tables one approach would be to make one colossal query of the confidential data that produces at once all the tables that the public will be able to see. Each number in each of these tables is a count, so the colossal query can be represented as a big matrix M applied to a huge vector $\boldsymbol{c}$ of the confidential data. DP would add noise to each count in $M\boldsymbol{c}$. But this may introduce more noise than is strictly required. A way to deal with this is known as the matrix mechanism [25, 19]. The public tables published by the Census are counts over discrete categories. The (confidential) data is a data base where each record has some attributes, and each attribute only takes on a finite set of values. These include age (from 0 to some upper bound), sex, Hispanicity, race (63 values), and so forth. An equivalent way of representing the data is as a (long) histogram, with one count for each possible combination of attributes. So there would be a count for 'male black-asian hispanics of age 37' and one for 'female white non-hispanics of age 12', and so forth. If these are arranged in some arbitrary order, we can think of the data base as a vector of counts $(x_1, x_2, \cdots, x_n)$. Then the result of a count query (e.g., 'male native-americans') is the inner product $w \cdot x$ where w is a vector of 0s and 1s of length n, with 1s exactly for those places in the histogram that count male native-Americans. This inner product is one of the counts in the publicly released tables. The set of queries that produce all these counts can be represented as the rows of a very large matrix W.

The idea of the matrix method is to answer all these queries (or this one giant query) in two stages. First answer a set of *strategy* queries in a differentially private way, and then combine the answers to these queries to get the queries we want ($W\boldsymbol{x}$). The strategy queries can be represented by some matrix A, one computes $\boldsymbol{m} = A\boldsymbol{x} + \Lambda$, where Λ is a vector of noise chosen so that the result is ε-DP. Then any post-processing of m does not affect privacy, so if $W = UA$, then $W\boldsymbol{x} = U\boldsymbol{m}$, which are the tables we want. One can attempt to find such an A that minimizes the mean error in the output. The process is illustrated graphically in

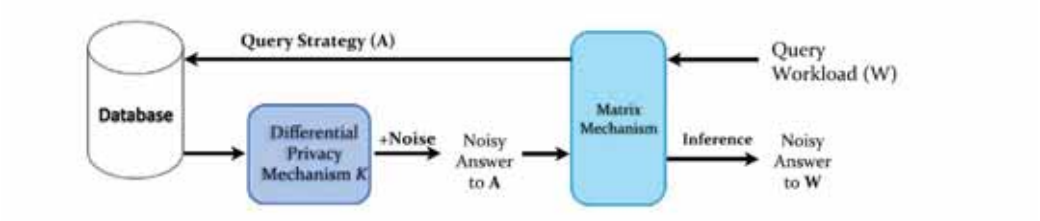


Figure 5-2: Process utilized by the matrix mechanism (from [25]).

Figure 5-2. This is a substantial computation described in the referenced papers.

5.2 Some Surprising Results in Applying Differential Privacy

The definition of DP does not immediately speak to the kinds of errors introduced. Nor does it guarantee that a query has a satisfactory (or any) DP algorithm. Below are presented some examples that indicate that one must be careful sometimes with the result of DP calculations to ensure statistical utility of the results.

5.2.1 Cumulative distribution functions

In [26] an example is given of how DP can affect common statistical measures. For example if we want to compute a cumulative distribution function (CDF) of incomes in some region we would count the number of income values less than some prescribed value and then divide by the total number of incomes to get a distribution. Under DP each time such a query is issued noise is added to the result. Depending on the level of noise injected the resulting CDF may become non-monotonic, something that is mathematically forbidden. Some results are shown in Figure 5-3 for a sample CFD under various values of ε. As ε is increased the generated CFD will converge to the smooth case without noise. The examples shown with a large amount of injected noise could not for example be reliably differenced to provide probabilities over small intervals. This is in fact the point - we cannot focus too clearly on the small scales. The issue identified here can be easily fixed by re-sorting the data so that a monotonic CFD results. The main

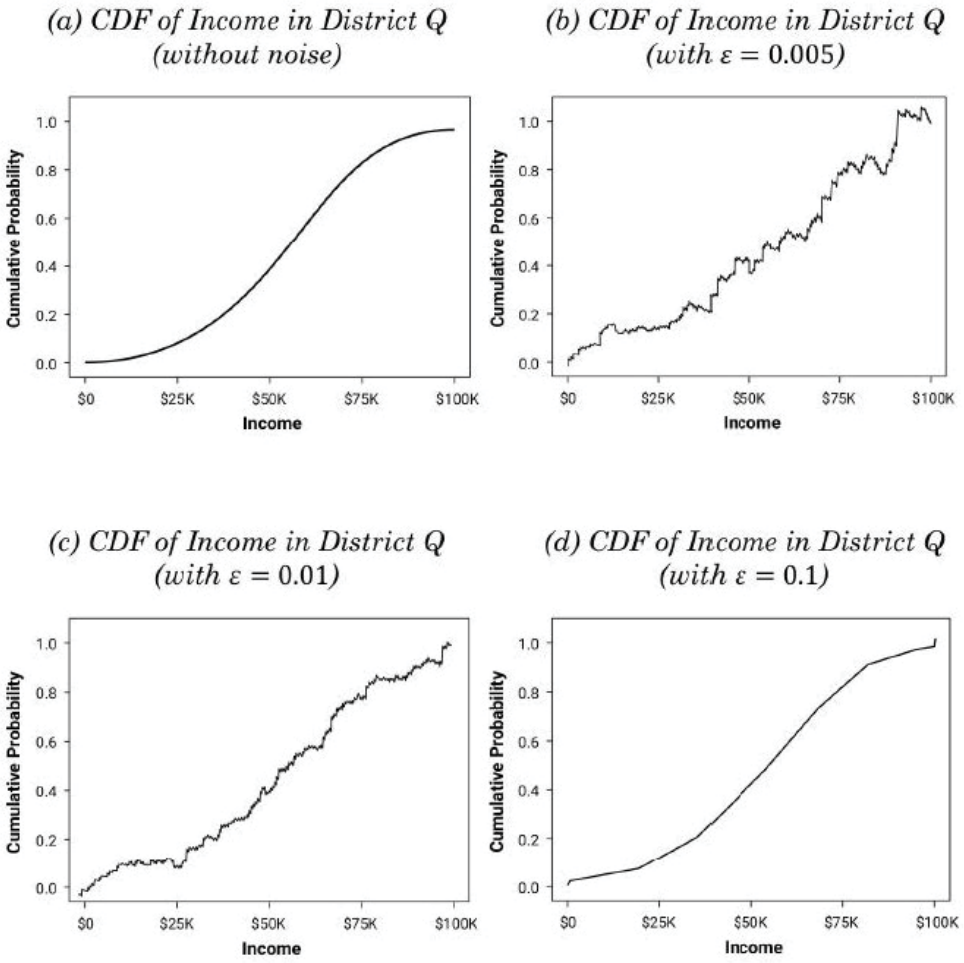


Figure 5-3: An example of a CDF of incomes under various values of the privacy loss parameter (from [26]).

point here is simply to point out possible issues with results published directly under DP.

5.2.2 Median

The examples of mechanisms so far involve additive noise, but the definition does not mention the type of noise. Consider a query that asks for the median. If the middle three elements in the larger database are 0.12, 0.14, 0.19, then if the size

of the database is odd, the median is 0.14, otherwise some tie-breaking algorithm would be needed. The smaller database is the result of removing one record from the larger database. If the number removed is no more than 0.12, the new exact median will be between 0.12 and 0.19. If the number removed is 0.19 or more, the same is true, and if 0.14 is removed, it is also true. So a privacy algorithm could choose any number between 0.12 and 0.19. Note that this algorithm decides what to do based on the data. It satisfies the intuition behind DP in that the result is independent of which record is removed from the database. However, it is *not* ε-DP for any ε. To see that, consider what the algorithm returns for the smaller database, if 0.12 were returned. Then the middle 3 might be 0.10, 0.14, 0.19, and the algorithm could return any value between 0.10 and 0.19. In particular there is a positive probability of returning a value in the interval [0.10, 0.12] for the smaller database, but that's impossible for the larger. So the ratio of probabilities in the definition of DP would be 0, which is impossible for any ε.

For the median, however, the sensitivity Δ is large. If the attribute takes on values between 0 and 1, and in the smaller database half of them are 0 and half of them are 1, then the median for the larger database is whatever value was removed, so $\Delta = 1/2$ (assuming that the algorithm chooses the midpoint for even sized databases). The Laplace mechanism doesn't look at the data, so it will add $\text{Lap}(1/2\varepsilon)$ noise. Answers that then fall outside [0,1] presumably would be truncated to be in range, so there is a positive probability of getting 0 or 1, which will almost always be silly and completely uninformative.

There is a similar story for any quantile, or the min, or the max, but the median is often used as a robust measure of location. Dwork and Lei [6] give a different algorithm that should be generally more satisfactory, but is data-dependent, and can fail (returning $\perp$ (null) in the language of computer science) on weird databases, such as the one in this example.

The decennial census data is just counts, so the peculiarities of medians are not directly relevant, but other statistical agencies and other statistical products

might not be so lucky.

5.2.3 Common mechanisms can give strange results for small n

Another mechanism is known as the random or uniform mechanism (UM). For a query that has a finite range, the random mechanism just chooses one uniformly; For example for the range of integers 0 through 10, choose a query response with probability 1/11. The random mechanism is ε-DP for any ε. If one were to propose a mechanism for a query associated with this finite collection of integers, it would seem undesirable for it to give the correct answer less frequently than the random mechanism does. That is, there may be many DP algorithms for the query, and it is unsatisfactory to chose one whose accuracy (meaning the chance of getting the right answer) is less than just choosing a result at random. For small n, both the truncated Laplace or Geometric mechanisms are unsatisfactory in this way.

There are various mechanisms for producing DP count data, The simplest way to think about these is to assume the data base has records with one sensitive field that has value 0 or 1. Suppose the query that counts the number of 1s needs to be protected. We know the answer is in the range [0, n], so the mechanism needs to produce a value in that range. The Range Restricted Geometric Mechanism (GM) produces

$$\min(n, \max(0, a+\delta))$$

where a is the true answer and δ is an integer chosen (at random) from a geometric distribution

$$(1-\alpha)^{|\delta|}/(1+\alpha)$$

where $\alpha = \exp(-\varepsilon)$ and ε is the parameter in differential privacy. Unfortunately, in this case, 0 and n will be over-represented. Worse, for most probability distributions on a, the actual count, if n is 2, the true answer of 1 is less likely than either of the incorrect answers 0 or 2. This is clearly a small n phenomenon,

but for small and modest-size n the usual mechanisms with various common loss functions give counter-intuitive results (cf. e.g. [4]).

Any mechanism for this problem is characterized by a (column) stochastic matrix P, where $P_{i,j}$ is $\Pr(i|j)$, the probability the mechanism returns i when the true result is j. P is an $(n+1)\times(n+1)$ matrix. The uniform or random mechanism (UM) has $P_{ij,} = 1/(n+1)$, that is, choose any answer at random. The set of all mechanisms can be defined by linear equations and inequalities. The only unobvious one, differential privacy, is expressed by

$$P_{i,j} \geq \alpha P_{i,j+1}, \quad P_{i,j+1} \geq \alpha P_{i,j}$$

for all i and j. The choice of a mechanism then comes down to minimizing some loss function over this polytope, preferably by linear programming. There are n^2 variables and a quadratic number of constraints.

Cormode's paper [4] notes that one can add a number of intuitively desirable constraints on the mechanism by adding linear constraints to this formulation. For instance, one might like the probability the mechanism returns the correct answer to be at least as large as the chance UM returns it, $P_{i,i} \geq 1/(n+1)$. Interchanging the values 0 and 1 in the statement of the problem converts a true answer a into $n-a$. One would expect the mechanism to be oblivious to this choice, which imposes a symmetry contstraint $P_{i,j} = P_{n-i,n-j}$. One would like the correct answer to be at least as probable as any other. The geometric mechanism (GM) satisfies these only for sufficiently large n, at least $2\alpha/(1-\alpha)$, which is roughly $2/\varepsilon$. If one adds the condition that answers closer to the true answer should be more likely than answers further away, then GM requires $\alpha < 1/2$.

For completeness, here is the explicitly fair mechanism of [4], which looks more complicated than it is, and satisfies their various sensible conditions:

$$P_{i,j} = \begin{cases} y\alpha^{|i-j|}, & \text{if } |i-j| < \min(j, nj) \\ y\alpha^{\lceil \frac{|i-j|+\min(j,n-j)}{2} \rceil} & \text{otherwise} \end{cases}$$

where

$$y = \frac{1-\alpha}{1+\alpha-2\alpha^{n/2+1}},$$

so the probability of returning the correct answer is a little larger than in the geometric mechanism, and the probabilities drop off more slowly with distance from the correct answer. The paper gives rules for choosing between this mechanism and GM.

5.2.4 Nearly equivalent queries with vastly different results

Suppose we have a database for which HIV-status is an attribute, with the values 0 or 1. The query might be "are more than half of the records 1?" One sensible way of answering this question using counts would be to ask for the size of the database n, and the number of ones, x, and look at the result. The returned values would have Laplacian or Geometric noise added to them, but unless the number of ones is very near 50%, the answer to the original question just pops out. A different computation, equivalent if exact results are returned, would be to ask if the median value of HIV-status is 0 or 1. As we have seen there is a positive chance of getting a meaningless answer regardless of how different the counts of zeros and ones. A more sensible query would be to ask for the average. The average is not a count query, but it has sensitivity $1/n$ for values between 0 and 1. So a DP query would answer with $\text{Lap}(1/n\varepsilon)$ noise added to the exact answer. This error drops rapidly with increasing n.

5.3 Invariants

The main promise of DP is to limit the knowledge that can be gained by adding or subtracting a record from a database. Informally if we make a small change in the input data the result of the output also undergoes a small change. That this is not always the case has been shown repeatedly through linkage attacks and database differencing. However, if certain results in a database must be openly

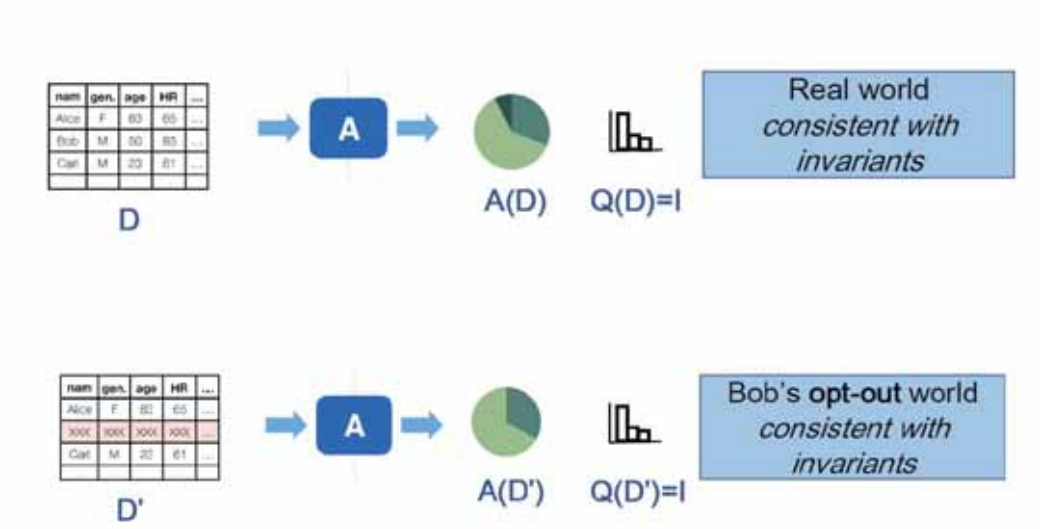


Figure 5-4: DP with invariants must be interpreted relative to a world in which respondents opt-out but consistent with invariants [21].

published without any protection then a small change in the input can have large consequence on the output if the output is directly tied to the small change.

An important example is the notion of an invariant. A simple example of an invariant relevant to the census is the need to publish an accurate count of the population of each state. For the 2020 Census, as in previous censuses, there are plans to publish state populations as exactly as possible and certainly without noise and so the state populations are invariants. In theory, releasing a true count is technically a complete violation of the DP guarantee. This is simply because removing one entry changes the population and so it is immediately obvious that a record has been removed even though we may not know which record.

As briefed to JASON by Prof. A Machanavajhala [21], it is possible to construct various scenarios where releasing an invariant could allow one to infer additional protected information regarding a record. There is to date no worst case characterization of privacy loss in this situation. At best, one can consider the incremental loss in releasing DP results in the presence of invariants. The situation is shown graphically in Figure 5-4. At present, it is not clear to what extent the

addition of invariants constitutes a vulnerability for Census data. As will be discussed below there are many more constraints that lead to invariants than just the population of the states. JASON does not know of a systematic approach to assess this except to perform a risk assessment by attempting to identify DP microdata as was orginally performed by Census in first identifying the existing vulnerability in the absence of noise. We discuss this further in Section 7.

5.4 Database Joins under Differential Privacy

In creating the various Census products such as SF1, the tables are produced through a join between two databases. One contains information about persons and the other about households. Queries such as the number of men living in a particular Census block requires only access to the person database while queries such as the number of occupied houses in a Census block requires only access to the household database. But if one wants to know how many children live in houses headed by a single man this requires a join of the two databases. Joins under DP can be problematic because one must examine the full consequences of removing a record in one table as it is linked to potentially multiple records in other tables. One way to address this is to create synthetic data as the Census is doing for both tables and then perform the join as usual. This however has been shown to produce high error in the results of queries essentially because too much noise is added for DP protection. A number of groups have researched this issue and provided possible solutions. The state of the art is a system called PrivSQL [15] which makes it possible to more efficiently produce tables via SQL commands while attempting to enforce a given privacy budget and while also attempting to optimize query accuracy. An architecture diagram for this system is shown in Figure 5-5. The system must generate a set of differentially private views for a set of preset queries. A sensitivity analysis must be performed and a set of protected synopses are then generated that can be publicly viewed. Census will perform the appropriate queries and create the protected tables using this

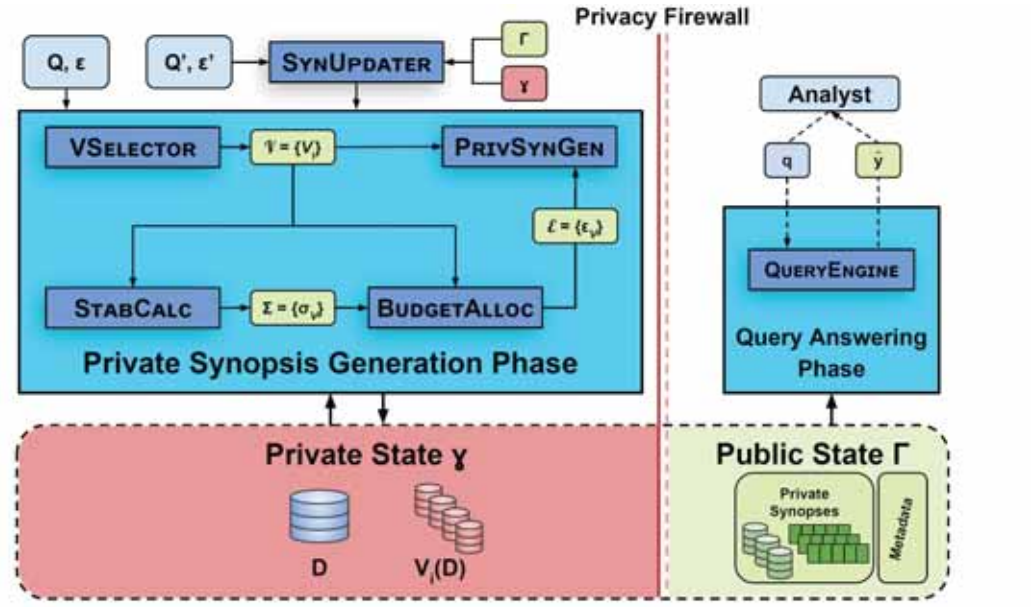


Figure 5-5: Architecture diagram for private SQL queries [15].

approach. Microdata associated with these tables will then be produced. This is at present work in progress, At the time Census briefed JASON their plan was to release a modified version of SF1 but tables requiring the linkage of data from person and housing records could not yet be constructed. It is expected that with further work using PrivSQL it should be possible to eventually produce many if not all of the traditional Census products.

5.5 The Dinur-Nissim Database under Differential Privacy

We provide here an example of the use of methods like DP as applied to queries of the Dinur-Nissim dataset. As discussed in Section 4.2 Dinur and Nissim made use of a simple database consisting of binary numbers to put forth what is now known as the Fundamental Law of Information Recovery, namely, that even in the presence of noise one can determine the contents of a private database by issuing and receiving the responses to too many queries. Here we illustrate that, despite the addition of noise, it is still possible to obtain meaningful statistical information from the database. We create a DN database as an array of randomly chosen

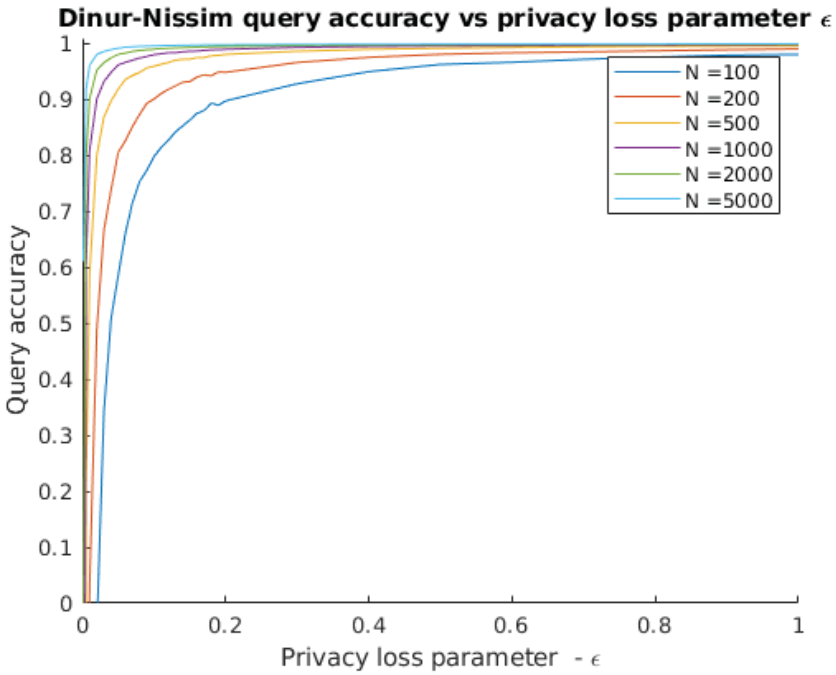


Figure 5-6: Accuracy of a sum query on the DN database. The values of N shown indicate the size of the database.

bits of size N bits. These could be the answer to a survey where the response is yes or no. We would like for example to know how many people responded yes to our survey. The result of our query is just the sum of the bits giving us the number of affirmative answers. For any query of this type issued we add a random amount of noise sampled from a Laplace distribution $\mathrm{Lap}(1/\varepsilon)$ with mean zero and variance $2/\varepsilon^2$. To measure the impact of the additional noise we calculate the query accuracy defined by

$$A = 1 - \frac{|\tilde{S} - S|}{S}$$

where $\tilde{S}$ is the noised sum and S is the sum in the absence of noise. A varies from 1 (no error) and then decreases towards zero and can become negative. Clearly, A of zero is of no utility. For each value of ε and N the number of bits we repeated the calculation 1000 times and reported the average A. The results are shown in Figure 5-6.

As can be seen, the accuracy of a query perturbed using the Laplace mech-

anism depends on the size of the data set. For the smallest dataset of size 100, a privacy loss value of $\varepsilon = 2$ degrades the query accuracy by about 15%. As N is increased the query accuracy improves and for $N = 5000$ the effect of the perturbation due to DP is imperceptible. In fact it would be smaller in this case than the statistical uncertainty associated with the query which varies as $1/\sqrt{N}$. For smaller values of ε the impact of the perturbation becomes more noticeable with the conclusion that smaller values of ε that provide increased privacy protection will not disturb statistical accuracy provided one deals with large datasets.

5.6 Multiple Query Vulnerability

As discussed in section 4 for the Dinur-Nissim dataset, it is still possible to recover the bits of the dataset provided enough queries are issued and optimization is used to get a "best fit" to the bit values. This works in our case even in the presence of arbitrarily large noise. The optimization technique, in our case least squares with constraints followed by rounding, can apparently return a result that converges to the true answer - the values of the bits in the dataset. We note that the residual norm of the optimization in this case will be very large, indicating that when the optimized result is used to compute the right hand side of the linear system representing the queries, the difference with the right hand side presented to the optimizer is very large. This is to be expected as we constrain the lower and upper bounds of the solution to be zero and one respectively. When we apply, for example, Laplace noise to the right hand side, we perturb it so that in some cases it would be impossible for a series of zeros and ones to sum to the indicated right hand side values. The larger is the noise amplitude, the more likely this is to occur. Nevertheless the optimizer will find solutions (effectively averaging out the applied noise) and as the number of random queries is increased the percentage of recovered bits increases.

To put this observation into the context of the Census vulnerability, we generate a Dinur-Nissim database consisting of 4000 randomly chosen bits. We then

generate a query matrix Q of size $N_Q \times n$ where n is the size of the database and N_Q is the number of issued random queries. In this case we set N_Q to be a multiple of the dataset size as this seemed more relevant to the issue faced by Census. That is, given a population, how many queries expressed as a multiple of the population suffice to infer the microdata. In the case of the Dinur-Nissim dataset, it is possible to ask this question even in the presence of noise and, empirically, while the number of queries required to determine the bits does increase with the size of the dataset, eventually, with high probability, all the bits can be recovered.

Given a query matrix and the dataset we compute the matrix-vector product and then set a value of the privacy loss parameter ε (in our case ranging from 0.01 to 1) and added to each component of the vector a random amount of noise sampled from the Laplace distribution. We then applied constrained least squares optimization and examined the fraction of bits recovered correctly. We assume that different bit locations are recovered correctly in computing the fraction recovered, but privacy concerns would certainly arise if the fraction of bits recovered exceeded 0.9. After some number of queries the algorithm succeeds in determining all the bits every time. A Matlab code performing this computation is included in Appendix B.

The results of our experiment are shown in Figure 5-7. Note that if one just guesses randomly, it is possible to recover 50% of the bits and so the minimum fraction of bits recovered is 0.5. The x-axis of the plot (labeled "Query multiple") indicates the number of queries scaled as a multiple of the size of the data set. In this case a multiple of 20 indicates 80000 random queries were made. The y axis indicates the privacy loss parameter. It can be seen that for example for $\varepsilon = 0.01$ and 4000 queries the results are not much better than random. But as the number of queries increases the fraction of bits recovered also increases. As the privacy parameter increases, and the number of query multiples increases eventually all the bits are recovered. This behavior is in line with the results of DP. Not only must one noise the data, one must also restrict the number of queries.

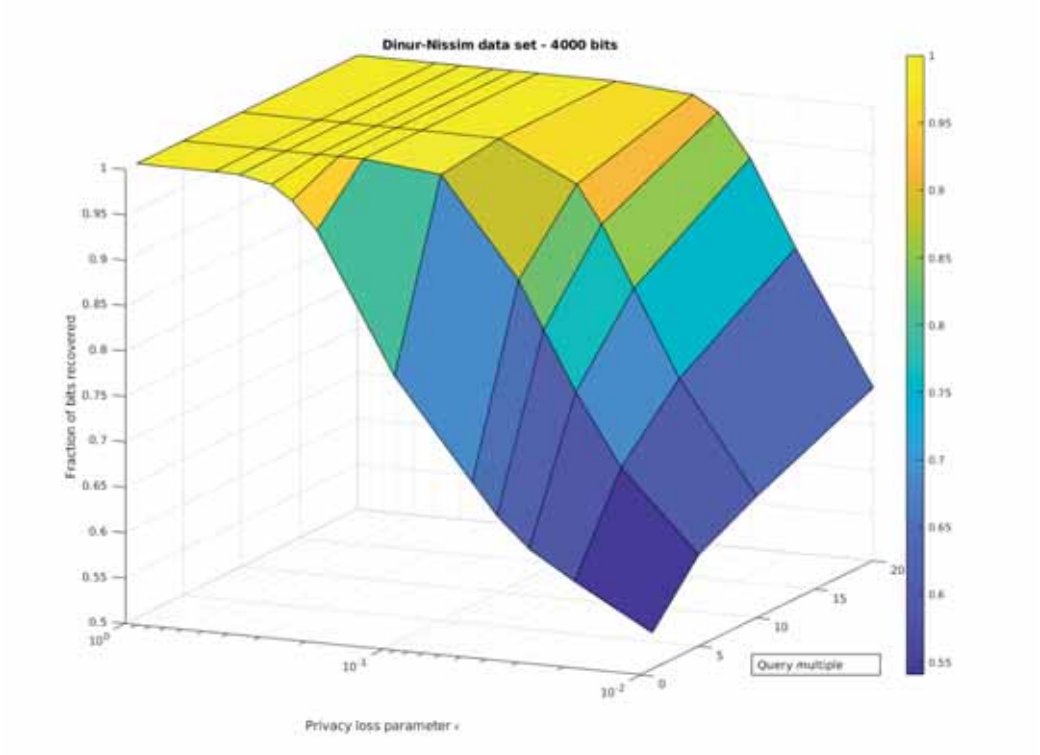


Figure 5-7: Fraction of bits recovered for the Dinur-Nissim database as a function of the privacy loss parameter and the number of multiples of the size of the database.

5.7 Disclosure Avoidance using Differential Privacy

The Census proposes to use an idea similar to that discussed above using the Dinur-Nissim database but applied to the much more complex microdata collected by the Census. As noted above, if one post-processes data that have been previously processed through an algorithm that satisfies the DP conditions, then the post-processed data will also satisfy the constraints of DP provided the original data are not accessed again during the post-processing.

If one creates the usual histograms as published by the Census (i.e. PL94, SF1, etc.) and then applies a DP mechanism to the results, then one could apply the same optimization technique used to demonstrate the Census vulnerability in Section 4 to produce microdata that are now themselves protected by DP. This

approach will create synthetic microdata upon which statistical queries can then be issued. We detail below the proposed approach following closely the briefing to JASON by Dan Kifer [14].

The approach Census will use has three phases

1. Select
2. Measure
3. Reconstruct

The microdata are first represented as a multidimensional histogram H. These are the tables that Census typically publishes. This histogram is then flattened into a column vector. A query on this histogram H is a linear function of the vector and can be represented by a query workload matrix Q. To acquire the answer to a prescribed set of queries we simply compute QH.

Selection phase In the selection phase a strategy matrix A is constructed for the purpose of optimizing the accuracy of various queries. A well chosen strategy matrix will minimize the sensitivity associated with the chosen queries by reducing the statistical variance of the queries. Algorithms for computing such a matrix are given in [20], but require some understanding of what the preferred query workload would be so that the appropriate set of queries is optimized for accuracy.

Measurement phase In this phase the query workload is performed with noise then added to the result. The amount of noise will depend on the sensitivity of the query and the chosen value of ε:

$$\tilde{Y} = AH + \text{Lap}\{\Delta_A/\varepsilon\}$$

where $\tilde{Y}$ is the DP response to the query and Δ_A is a norm measuring the sensitivity of the strategy matrix A.

Reconstruct The final step is to estimate QH from the vector $\tilde{Y}$. This requires undoing the multiplication by the strategy matrix:

$$QH = QA^{+}\tilde{Y}$$

As the strategy matrix may not be square, the Moore-Penrose pesudo-inverse is used to compute H and then QH.

The measurement phase consumes the privacy budget. Once this is accomplished the results could in principle be released to the public. The reconstruction phase will not re-access the private data and hence does not require additional privacy budget. The cleverness of this idea is that the final product can even be in the form of microdata which can then be reprocessed by users of the Census data. What is less clear however, is the accuracy of queries that have not been optimized using the High Dimensional Matrix Method and whether the results of those queries will have an acceptable statistical utility. This will be discussed further in Section 6.

While the steps of this procedure are easily described, the computational aspects of doing this for the census pose significant challenges. Recall that for the country Census publishes billions of queries and so the histogram will have billions of cells. The query matrix could be as large as the square of the histogram size depending on what measurements are to be reported. Choosing a strategy matrix based on the potential query workload is not feasible. The reconstruction is also going to entail an enormous computational cost as a a result of the matrix sizes. Finally, the result of the multiplication by the Moore-Penrose inverse will lead to non-integer results. If we wish to convert these to sensible microdata a second phase will be required in which the results of the first phase will have to be converted to integers. Once this is done the optimization approach taken by Census to reconstruct the microdata can be used to create differentially private microdata.

The solution to the challenges discussed above are to break the problem up into pieces and then perform the DP reconstruction on each piece. The first

attempt to do this was a “Bottom Up” approach in which the select-measure-reconstruct approach was applied to each Census block and then converted to microdata. This has the advantage that the operations are all independent for each block and the privacy budget is simple - one value of ε can be assigned to each block. The privacy cost does not depend on the number of blocks as each of these is processed independently of the others. It also has the advantage that the counts at various levels of the Census hierarchy are consistent. However, the injection of the DP noise adds up as the data are combined to form results for block groups, tracts, etc. A county in a populous region that contains many blocks will have an error proportional to the number of blocks. The “Bottom Up" approach is easy to conceptualize but it doesn’t use the privacy budget efficiently.

Instead, Census will use a “Top-Down” approach. The privacy budget is split into six parts: national, state, county, tract, block group and block. A national histogram $\tilde{H}^0$ is then created using the select measure and reconstruct algorithm outlined above. This involves the population of the US but the number of queries is now manageable as the queries are not specified over geographic levels finer than the nation. Once this protected histogram is in place the same process can then be applied for the states using the privacy budget allocated for states. These histograms are constrained so that they are consistent with national totals. This process is then followed down to the county, block group and finally the block level. Once a protected histogram with non-negative integer entries is created it can then be transformed to microdata using the optimization approach Census used to determine the reconstruction vulnerability as discussed in Section 4. The Top-Down approach has the advantage that it can be performed in parallel and the selection of queries can be optimized at each level making it possible to use the privacy budget more efficiently. It also has the advantage that it enforces any sparsity associated with 0 populations at various levels (for example someone over 100 who indicates they are a member of five racial categories). These are known as structural zeros.

In producing an appropriate histogram that can be turned into microdata two

optimizations are performed. The first is a least squares optimization which effects the Moore-Penrose inverse subject to various constraints that the histogram being determined must be consistent with the parent histogram. For example the total population of the states must sum to the population of the country. The result of this optimization leads to fractional entries and so the second step is to perform an optimization that assigns integer values to the histogram cells such that the entries are non-negative integers that are rounded values of the fractional results and that sum to the same totals consistent with the parent histograms. This "rounding" step is performed using the Gurobi solver [12].

A complication in executing the TopDown algorithm is the need to publish some data without protection. These correspond to the invariants discussed in Section 5.3. Census plans to provide accurate counts of the population of each of the 50 states, DC and Puerto Rico to support apportionment of Congressional representatives. It might also be desirable to report correct population down to the census block.

But in addition, there are other constraints and so it would be desirable to be consistent with these. For example, the number of occupied group quarters and housing units in each census block is public information as a result of a program called Local Update of Census Addresses (LUCA). This program is used by Census to update the Master Address File (MAF) used to distribute census surveys. The addresses themselves are protected under Title 13 but the number of group quarters is publicly released. As a result, if a census block were to have an occupied jail then the TopDown algorithm must assign at least one person to that jail. As another example, the number of householders in a block should be at least the number of households [14]. There are other data-independent constraints. For example, if a household has only one person in it then that person is presumably the householder.

Census has proposed a partial solution to this problem by casting the constraints as a series of network flows that can then be appended to both the least

squares and rounding optimizations described above [14]. This work is still experimental at the time of this writing and will be further evaluated.

The enforcement of invariants such as national and state populations presents no issues in terms of the DP computation. Neither does the enforcement of structural zeroes such as there cannot be any males in a dormitory that is all female. But the constraints that are independent of the data such as the fact that a grandparent must be older than the children in a household creates issues of infeasibility as the optimization recurses down the Census geographic hierarchy. If such implied constraints are ignored there is the possibility that for example assignments at the block group level are not consistent when extended to the higher Census tract level. When this happens it is called a "failed solve"and Census then applies a "failsafe" optimization. The constraints impeding solution are relaxed and the optimizer finds the closest feasible solution meaning a violation of the exact constraint will be allowed. The assignments at the higher geographic level (for example the county level of optimization at the tract level fails) are then modified to maintain hierarchical consistency. The overall impact of the use of the failsafe on the utility of the protected Census data is still not fully understood and is an area of ongoing research. One approach that would avoid this difficulty is to not insist on hierarchical consistency at the finer geographic levels, in particular census blocks. For example providing the correct population in each block might not be enforced as a constraint. This however may have implications for the use of census data in the redistricting process, an issue we discuss in Section 6.

The new disclosure avoidance scheme will now look as in Figure 5-8. It is expected that Census will still perform the usual imputations associated with households and general quarters for which Census enumerators cannot obtain information but, at present, no household swapping will be performed. Instead the Census will apply the TopDown algorithm and then create a set of noised tabular summaries and also, for the first time, the synthetic microdata associated with the summaries.

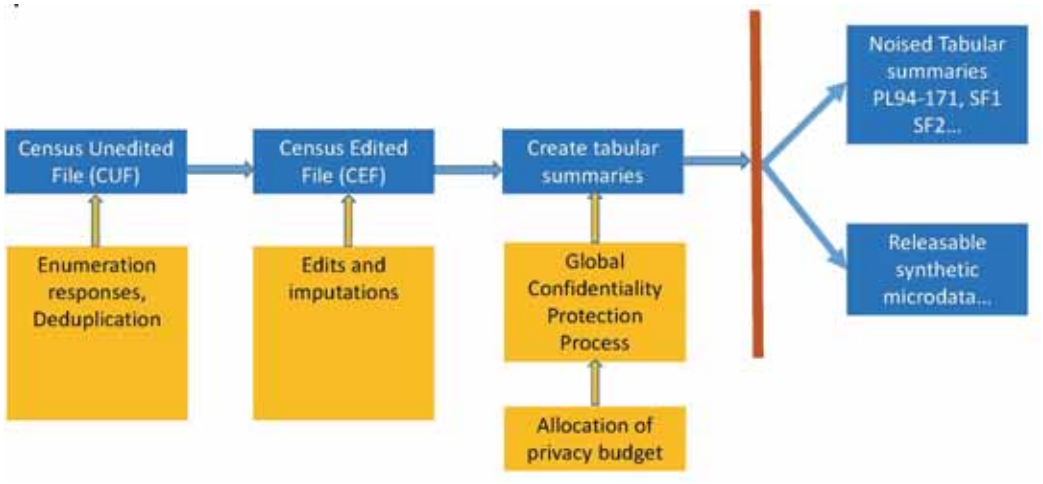


Figure 5-8: A graphical representation of the proposed DAS using the TopDown DP algorithm.

The proposed disclosure avoidance system using DP has been implemented in Python and is publicly available [34]. Work continues to improve query accuracy and enforce invariants and implied constraints. Census is to be commended for making this software available to the community so that it can be examined in detail and inform users on the details of the application of DP to census data.

6 ASSESSING THE ACCURACY-PRIVACY TRADE-OFF

In this section we examine the results of some of the early applications of the new Census DAS on census data. As mentioned in Section 5 Census has publicly released the DAS software. To further aid users, it has processed census data from 1940 and produced synthetic microdata. It has also released some preliminary assessments of query accuracy for the 2010 census data. We discuss these results here with an emphasis on the trade-off between query accuracy and the level of privacy protection.

6.1 Census Analysis of 2010 Census Data

Census has applied the proposed DAS using DP to the 2010 census data. The advantage here is that the schema for the 2010 census largely overlap with the schema for the forthcoming 2020 census. But a disadvantage is that this data is not yet publicly available. By law census data can only be publicly released no earlier than 72 years after a census is taken so the latest data available to the public is the 1940 census. We are able to provide only a limited view of the results of the Census analyses on 2010 data as most of these are not yet available for release and are still protected under Title 13. JASON did have access to these results but the assessment provided here can only describe them qualitatively.

As briefed to JASON by P. LeClerc [16], Census has executed the TopDown algorithm on a histogram from the Census Edited File H_{CEF} to produce a noised histogram of privatized results H_{DAS}. The experiments were performed for the PL94-CVAP product that has 4032 entries representing a shape of $8 \times 2 \times 2 \times 63 \times 2$. Recall that this product is used to examine voting districts to ensure adherence to the Voting Rights Act and includes the following pieces of information:

- 8 group quarters-housing units levels,

- 2 voting age levels,
- 2 Hispanic levels,
- 63 OMB race combinations,
- 2 Citizenship levels.

For each state one can create such a histogram and examine it at various geographic levels: state, county, tract, block group and block. For each geographic level (geolevel) γ, Census executed 25 trials of the DAS, averaged over the results, and reported a number of metrics. We will consider here only one of them:

$$\text{TVD}_\gamma = 1 - \frac{L^1(H_{DAS,\gamma}, H_{CEF,\gamma})}{2\text{POP}_\gamma}.$$

This can be thought of as a type of accuracy metric using the L_1 norm or sum of the magnitudes of the distance between the DAS and CES entries. This is similar in some respects to the Dinur-Nissim query accuracy metric discussed in Section 5.5. If the DAS and CEF histograms were to agree across all components at a given geographic hierarchy level γ, the TVD value would be exactly 1. The possible difference between the values is normalized by twice the population, but this does not provide an absolute lower bound on the TVD metric and it can become negative depending on how much noise is infused into the histogram values.

As of the date of this report, Census has publicly released TVD metrics for the state of New Mexico [30]. These indicate query accuracy vs. privacy loss for actual Census data and may be reflective of the results of the future 2020 Census. In Figure 6-1, the TVD metric as a function of ε is plotted at the state, county, tract group, tract, block group and block for the state population. As ε increases from 0, the TVD metric will tend to one indicating that as ε increases less noise is injected into the histograms until at sufficiently large ε the DAS and CEF results agree in this norm. As can be seen, for geolevels with large populations (e.g. counties, tracts and even block groups) the TVD metric for population is close to one for values of ε as small as 1/2. At even lower levels of ε we see the same

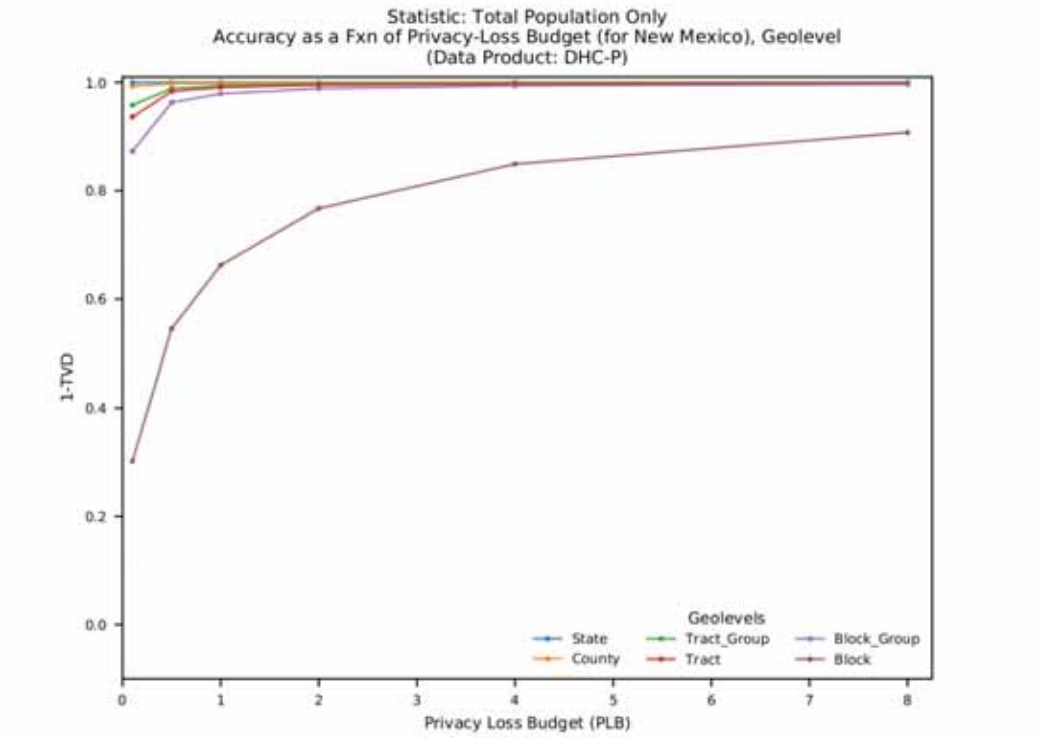


Figure 6-1: A plot of the TVD metric for total population for various geolevels as a function of privacy loss parameter for the state of New Mexico [30].

type of degradation of query accuracy as in the Dinur-Nissim example. Because we cannot tie TVD to a measure of statistical accuracy we cannot comment on whether such degradation of accuracy would or would not be acceptable from that point of view. At the block level, because populations are typically much smaller than block groups the degradation is noticeable and even at $\varepsilon = 4$ we still have TVD ≈ 0.8.

In Figure 6-2 we show again the TVD metric but this time for a subhistogram looking only at those entries associated with race and Hispanic origin. Typically the counts here will be smaller particularly as we examine the finest block level and so the TVD metric deviates further from 1 than shown in Figure 6-1 as the privacy loss budget is decreased.

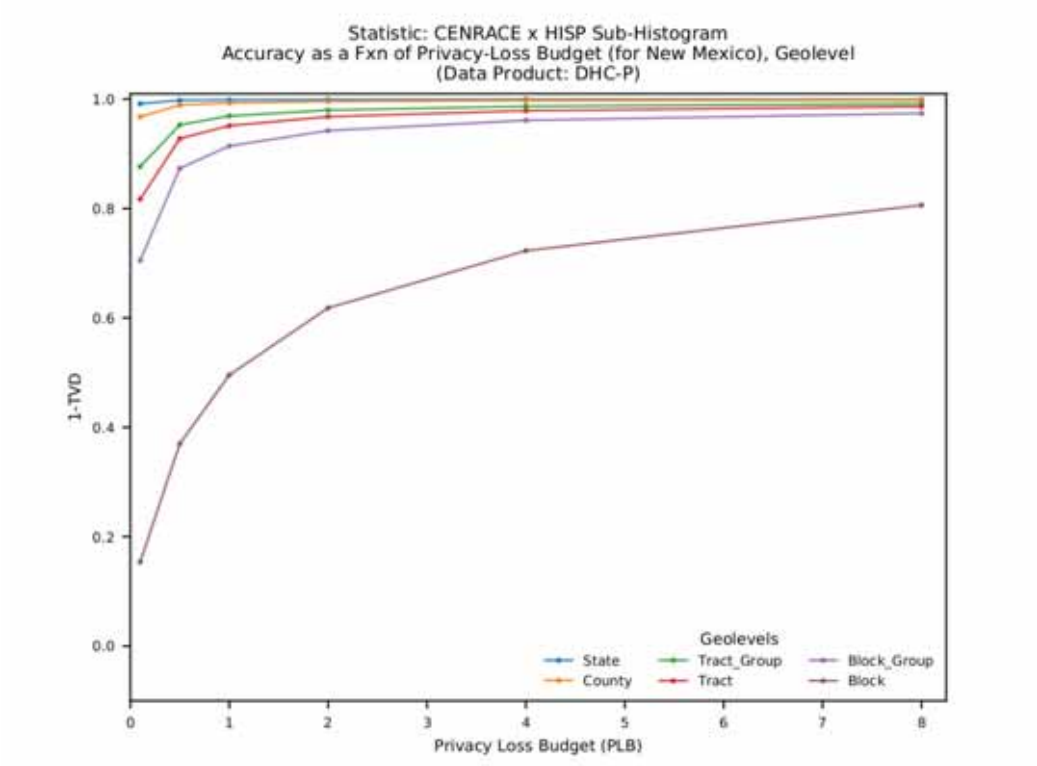


Figure 6-2: A plot of the TVD metric for race and Hispanic origin for various geolevels as a function of privacy loss parameter for the state of New Mexico [30].

The TVD metric provides some insight into the degradation of query accuracy as the privacy loss budget is decreased, but it suffers from being a coarse measure of accuracy as it sums over the entries at a given geolevel and so does not provide a view of the variance of the individual differences. For example, it would be useful to see the distribution of TVD measure block by block. A more detailed assessment in terms of microdata but for the older 1940 Census is discussed in the next section.

6.2 IPUMS Analysis of 1940 Census Data under the Census DAS

IPUMS (Integrated Public Use Microdata Series) is an organization under the University of Minnesota Population Center providing census and survey data from a

variety of countries. It is the world's largest repository of census microdata. JASON was briefed by Dave van Riper of IPUMS [36] (cf. also [37]) who examined in detail the application of the Census DAS to the 1940 Census microdata. We note that JASON has not verified this work but we discuss it here to give examples of the differences between counts associated with the DAS processed synthetic microdata and the true census microdata. As discussed in Section 4, we expect more dispersion as we descend to finer geographic regions. At the time of van Riper's briefing he had performed comparisons for Minnesota census data. Since then, he has also performed analyses for the entire US and it is this data that we discuss here.

It should be noted that the geographical hierarchy for the 1940 census was different than that used today. The finest level of geographic resolution is what was then called an enumeration district. Enumeration districts are roughly comparable to census block groups on the geographic spine and also similar in some ways to what Census terms "places". The median population for enumeration districts was about 1000 people. The median population for census places in 1940 was about 800 people.

As indicated in Section 5, Census has publicly released differentially private microdata for the 1940 census. Microdata files were generated for the entire country for eight different values of the privacy loss parameter ε : 0.25, 0.5, 0.75, 1.0, 2.0 4.0, 6.0, 8.0. Four runs of the DAS were provide at each value of ε. The microdata made available are those of the PL94-CVAP Census product and include whether a respondent is of voting age, Hispanic origin and Race as well as household and group quarters type at four geographic levels: national, state, county and enumeration district. IPUMS did not run the Census DAS to generate synthetic microdata. Instead it analyzed those results generated by Census to compare against unfiltered microdata that constitute ground truth. The source code for the DAS system [34] is configurable so that one can allocate fractions of the total privacy budget over the various geographic levels and tables. In this case the budget is allocated evenly over geographic levels. Each level of the hierar-

chy receives a quarter of the total privacy budget. Allocations must also be made for the various tables that are produced and then subsequently noised by the DP algorithm. In this case Census chose the following fractions:

- Voting age by Hispanic Origin by Race: 0.675
- Household group quarters type: 0.225
- Full cross of all variables: 0.1

The fraction of the total privacy budget to be allocated for each level and for each table is then the product of the geolevel allocation times the table fractions. For a given total privacy loss budget ε it is these fractions that are used to provide the noise levels for each individual table at a given geographic level. For example if the total privacy budget were 0.25 then the privacy budget for each histogram will look as shown in Table 6-3. The table shows the effective values of ε but also the level of dispersion for an equivalent Laplace distribution. These dispersion levels will affect various tables differently. A table associated with large counts will not be significantly affected by an ε corresponding to a dispersion of 300 but a table at the enumeration district level could be significantly affected.

Box plots of the distribution of populations across all US counties in 1940 are shown in Figure 6-3 for all the values of ε used in the Census runs of the DAS. The distribution as computed by IPUMS from the true 1940 microdata is shown at the left of the Figure. As can be seen, as ε increases the box plots converge to the IPUMS result. For the lowest value of ε used, differences can be seen for populations of 100 or more. By and large, the box plots are quite similar across the various values of ε. More insight into the effect of the DAS at the finer geolevels can be seen in Figure 6-4 where box plots for the differences between the DAS and IPUMS population estimates are shown. The orange box plots represent counties and the teal plots represent enumeration districts. Again as ε increases we see the differences reduce. But at lower values of ε differences on the order of several hundred people appear when we look at various outliers. It should be noted

Geography level	Table	ϵ for Table at level	Noise dispersion
Nation	Vot–Hisp-Race	0.042	47.4
Nation	HouseholdGenQuart	0.014	142
Nation	Detailed	0.006	320
State	Vot –Hisp-Race	0.042	47.4
State	HouseholdGenQuart	0.014	142
State	Detailed	0.006	320
County	Vot –Hisp-Race	0.042	47.4
County	HouseholdGenQuart	0.014	142
County	Detailed	0.006	320
Enum Dist	Vot–Hisp-Race	0.042	47.4
Enum Dist	HouseholdGenQuart	0.014	142
Enum Dist	Detailed	0.006	320

Table 6-3: Values of the privacy budget allocated to the various geolevels and tables by the Census DAS system for the 1940 Census data [36]. The noise dispersion is listed here to give some notion of the variance of the noise applied to the data. In this case the value $\varepsilon = 0.25$ is used [36]

.

that the box plots are not normalized and that the teal box plots for enumeration districts are smaller simply by virtue of representing smaller populations.

Van Riper has also computed how the populations of counties compare in detail in Figure 6-5. The Figure plots the IPUMS value for a county population vs. the DAS value. The level of agreement is measured by how closely the two values would lie to the 45° line indicating equality. As can be seen the county populations align well at all values of ε. In contrast, for enumeration districts we see in Figure 6-6 more dispersion. This is most observable as ε becomes smaller. Note that because the DAS does not allow negative population there is a pile-up as population size decreases. Such results are to be expected as one focuses on finer geolevels and smaller populations.

The same analysis has been performed for population under 18 across all US counties for the 1940 Census. These are shown in Figure 6-7. This too looks quite

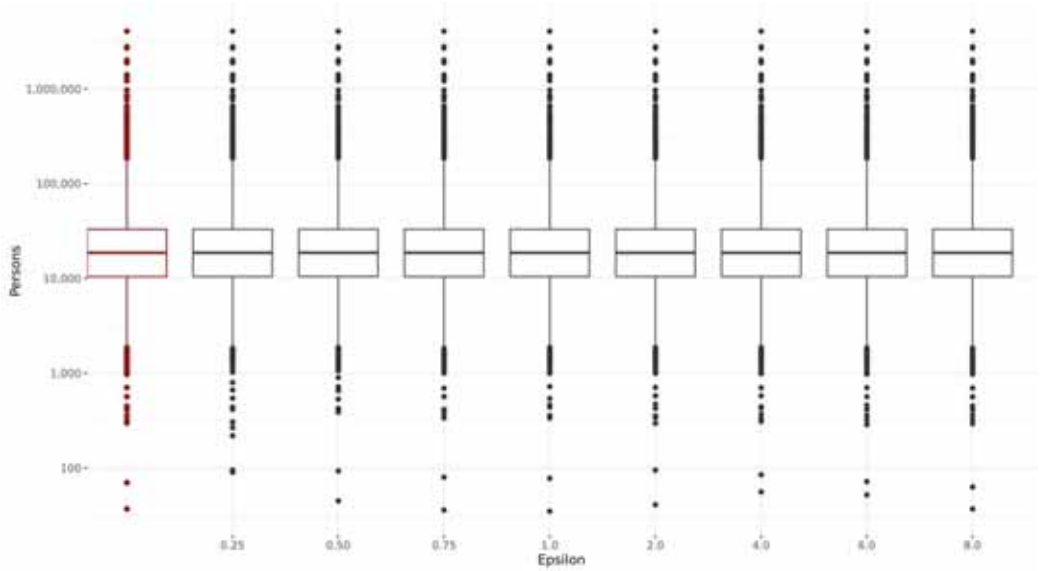


Figure 6-3: Box plots for the distribution of total US population in 1940 under different values of the privacy loss parameter [36].

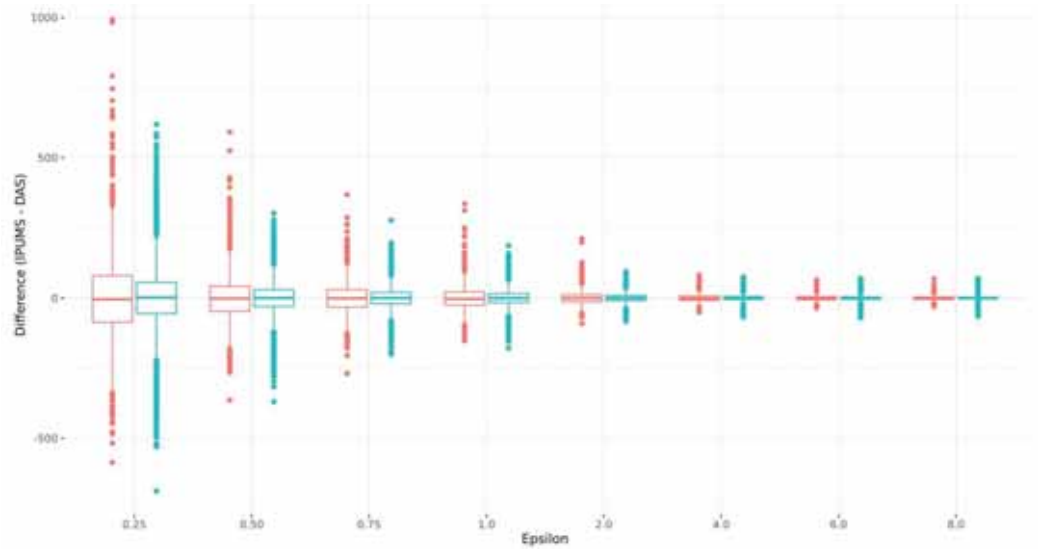


Figure 6-4: Box plots for the differences between IPUMS and Census DAS for total population counts under different values of the privacy loss parameter [36].

similar to population estimates with some issues seen for counties with smaller populations at lower values of ε. The corresponding results for enumeration districts are shown in Figure 6-8. Because we are now focusing on a subgroup of the population for enumeration districts there is yet more dispersion in the results. But

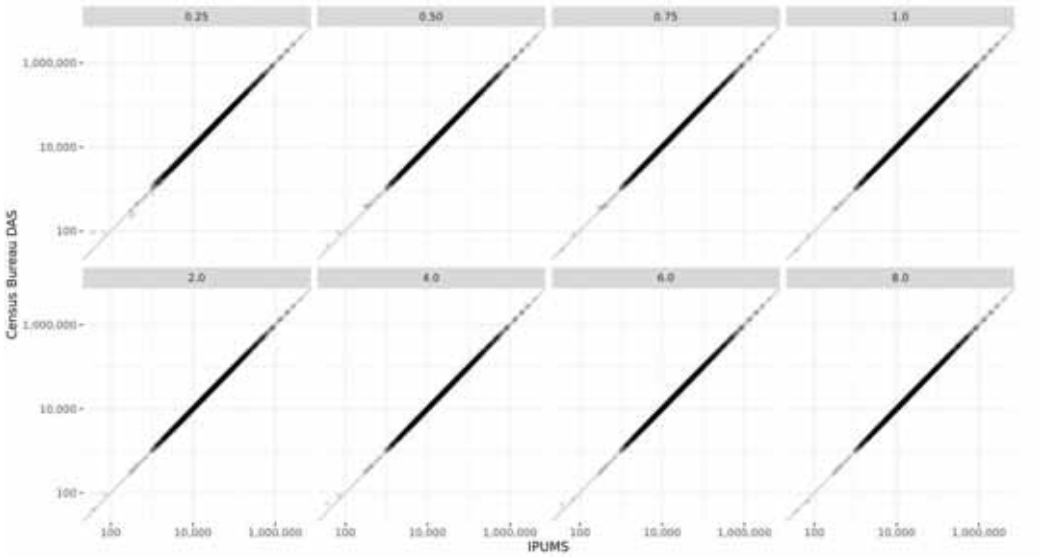


Figure 6-5: Total population for US counties under differing levels of the privacy loss parameter [36].

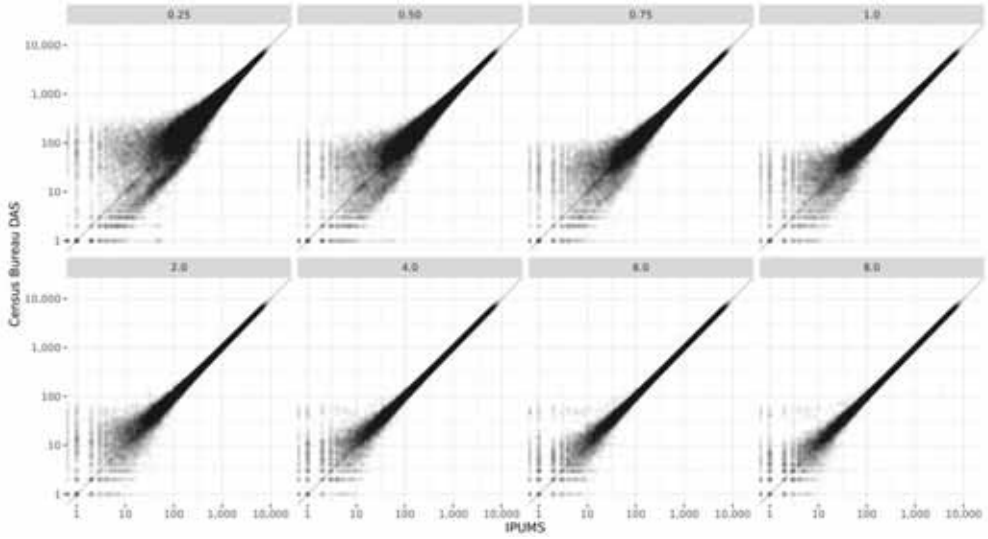


Figure 6-6: Total population for US enumeration districts under differing levels of the privacy loss parameter [36].

perhaps of some concern is that in some enumeration districts the DAS indicates a large number of people under 18 when there are in fact very few. There are some enumeration districts with 50 or more people where this particular application of the DAS (with values of ε of 0.25, 0.5 and even in some cases 1.0) indicates that 100% of the population is under 18, an observation that could have implications

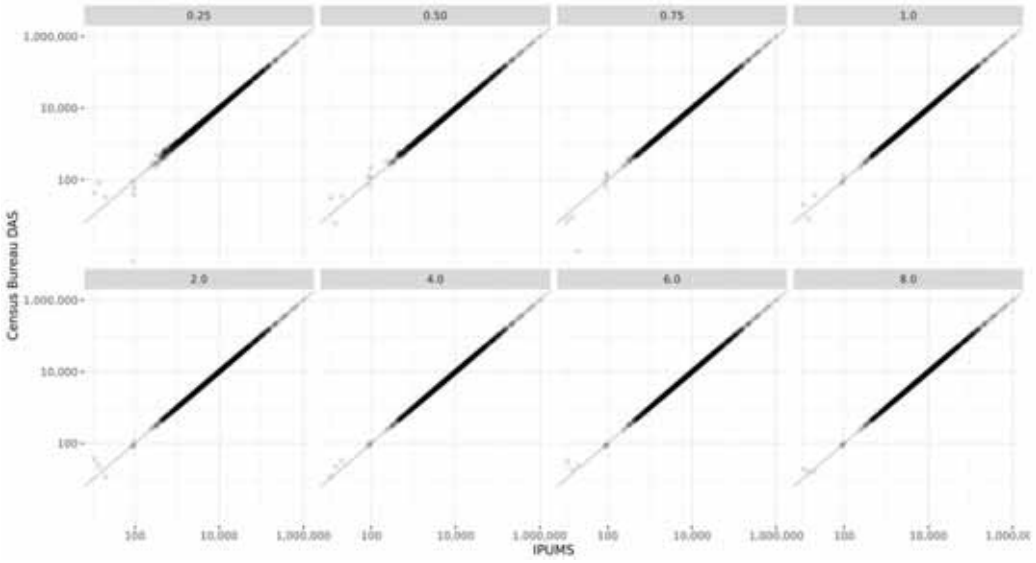


Figure 6-7: Total population under 18 for US counties under differing levels of the privacy loss parameter [36]

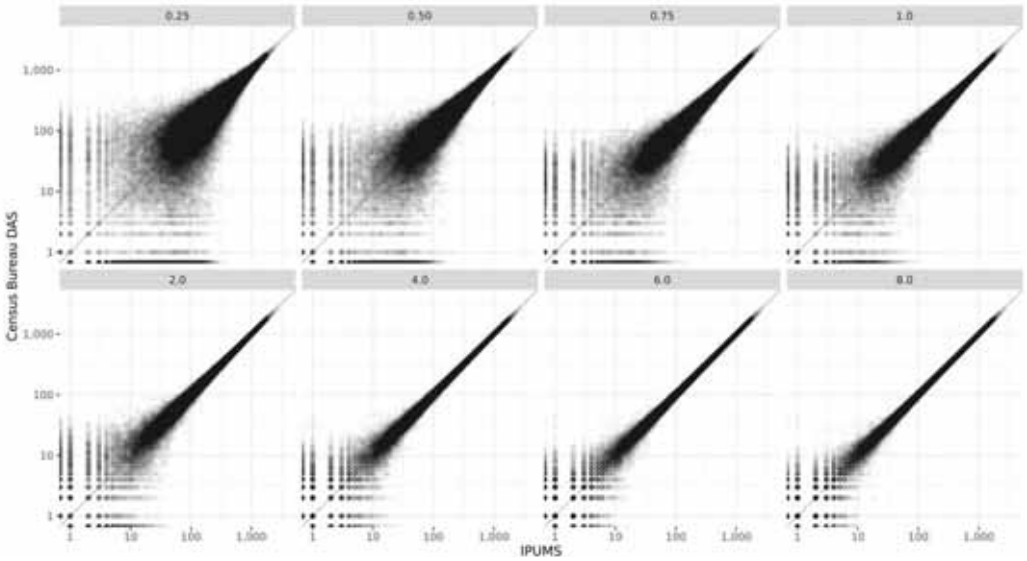


Figure 6-8: Total population under 18 for enumeration districts under differing levels of the privacy loss parameter [36]

for assessments of voting age population, a component of the information needed for the PL94 publication.

Several points should be emphasized in examining the current application of the DAS:

- The DAS does not unduly perturb statistics at the national, state and even largely at the county level at all the values of ε considered.
- The dispersion seen in the IPUMS-DAS comparison for enumeration districts is to be expected at lower values of ε. The DAS is after all meant to protect small populations.
- The application of the DAS will degrade the utility of various statistics. This degradation will increase as one further restricts the population by characteristics such as race, voting age, etc. This illustrates a trade-off inherent in the use of DP among privacy, accuracy and granularity of queries. The requirements for accuracy will need to be determined in the future through consultation with external users of the data. We discuss this trade-off further in Section 7.
- The allocation of the privacy budget can be modified depending on the accuracy requirements. For example it would be possible to allow for larger privacy loss parameters for some tables and less for others provided the total privacy budget is conserved.
- The current version of the DAS is a demonstration product. For example, at the time of this writing, the implementation presented here does not benefit from the improved accuracy of the high dimensional matrix method. Nor do the products contain all the invariants and constraints that the Census bureau has identified. Work is in progress to improve query accuracy to the extent possible. As these improvements are made it will be important to continue to reevaluate the performance of the DAS against ground truth.

7 MANAGING THE TRADE-OFF OF ACCURACY, GRANULARITY AND PRIVACY

Published census tabulations must balance inconsistent desiderata. They should be accurate (i.e., published counts should be the sums of the underlying micro-data). But tabulations should also be appropriately granular (i.e., have a high level of detail such as block, gender, age, race/ethnicity, etc. But, as has been discussed, pushing granularity to the extreme can create small (or even singleton) counts in table entries (particularly in small blocks), thereby eroding privacy. Of course, privacy could be enhanced and granularity preserved by relaxing the accuracy requirement (as embodied in DP or swapping schemes). Alternatively, privacy could be enhanced and accuracy preserved by reducing granularity. The situation can be illustrated by the "disclosure triangle", where the balance among the three competing considerations of privacy, accuracy, and granularity varies across the interior as shown in Figure 7-1.

No compromise will be perfect. In this section, we discuss some aspects of managing this trade-off.

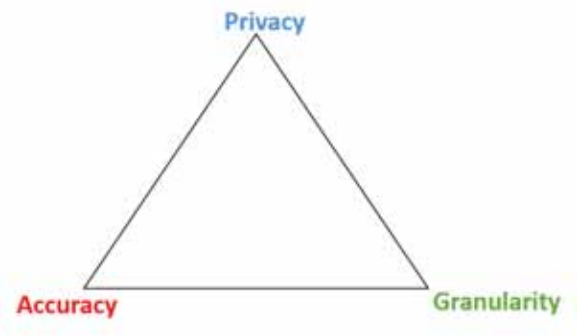


Figure 7-1: Census must balance, accuracy, granularity and privacy in its publications. It is not possible to achieve all three simultaneously.

7.1 Risk Assessment

The use of DP is clearly promising as a way to protect census data, but it is important to recall the original motivation for its use. Its proposed use was primarily motivated by the 17% re-identification rate assessed by Census using the 2010 tables, and thus the degree to which DP prevents re-identification needs to be similarly explored. Technically, differential privacy as pointed out by Reiter [28] is a guarantee

> "on the incremental disclosure risks of participating (in a survey) over whatever disclosure risks the data subjects face even if they do not participate (in the survey)".

It does not provide an assessment of disclosure risk in and of itself. It is also not one methodology. A number of algorithms can be applied and must be implemented correctly. In the case of its use for the census there are clearly complications like invariants, implied constraints etc. that will require further work and assessment. For these reasons, explicit quantification of the risk of re-identification is still required. The choice of ε should be informed by calculations of the risk of re-identification using the methods developed by Census and linking with current commercially-available data but applied to microdata as processed through DP. JASON understands that this will be significantly more difficult than the original analysis that led to the re-identification of the 2010 Census data vulnerability. This is because the matching of the microdata in the absence of noise to commercial data was aided by the availability of the geographic location. The synthetic data generated by DP algorithms will not have this feature and so matching to commercial data bases will have to be performed using probabilistic record linkage (cf. for example [9]). A very useful property of DP here is that such linkage can be attempted at various values of ε. At very high values of ε we expect to recover the noise-free values and so we would also verify the previously assessed re-identification level of 17% against commercial marketing databases. But as ε

is decreased this re-identification rate must degrade. An open question is at what value of ε would it degrade to a value sufficiently low so as to be administratively acceptable? While no official value of such a lower bound has ever been provided (nor would we expect one to be) presentations from Census have indicated that the re-identification rate of 17% was viewed as something like four orders of magnitude higher than previously assessed [27].

The fact that methods of data science will improve and commercially available data will become more comprehensive over time does not obviate the need for an analysis that can inform the current decision. Knowing the outcomes based on current data can help to support a choice of ε. Once some assessment of an appropriate "upper bound" for ε based on disclosure risk is in hand, further considerations regarding statistical accuracy for future queries on the data can be made in ultimately deciding the level of noise to be applied to the 2020 data.

7.2 Engaging the User Community

Analyses of aggregate data involving large populations will be minimally impacted by DP. Impacts will increase as one focuses on finer levels of geography or other demographic measures. We emphasize that this is precisely the desired impact of DP because individuals within a smaller group will be more identifiable, and thus it is precisely this "blurring" from DP that protects the privacy of these individuals. This aspect of DP needs to be effectively communicated to future users of Census data.

The challenge is to better quantify the balance of privacy protection and data utility for smaller groups. There are multiple communities with a deep interest in the accuracy-privacy-granularity tradeoff:

State governments and redistricting commissions These bodies are responsible for the drawing of Congressional and State legislative districts. PL94-171 requires the Census to provide to these bodies an opportunity to identify

the geographic areas relevant to redistricting and to then deliver tabulations of the population as well as race, race for population 18 and over (voting age), Hispanicity and Hispanicity for those 18 and over, occupancy status and, in 2020, group quarters population by group quarters type.

Local governments Local governments use census data for redistricting as well as to inform assessments of public health, safety, and emergency preparedness for the residents.

Residents Residents use census data to support community initiatives and to decide where to live, learn, work and play.

Social scientists and economists Census data forms a foundation for demographic studies as well as economic research.

Census has to some extent reached out to these communities through a July 2018 Federal Register Notice as well as several academic conferences [23]. The feedback received by Census emphasized several aspects:

- There was little understanding as to the need for application of Differential Privacy
- Users were vocal about the need to maintain block level data so that custom geographies could be constructed.
- Concerns were voiced about the potential loss of information for small geographic areas.

Clearly more work is needed and Census should participate actively in various fora, working with the community to characterize the scales and types of queries that will and will not be substantially impacted at different values of ε. For example, opportunities for stakeholders to assess accuracy of queries on 2010 census data made available at various levels of protection would go a long way towards helping users assess the impact of DP on future analyses. In general it

will be necessary to engage and educate the various communities of stakeholders so that they can fully understand the implications (and the need for) DP. These engagements should be two-way conversations so that the Census Bureau can understand the breadth of requirements for census data, and stakeholders can in turn more fully appreciate the need for confidentiality protection in the present era of "big data", and perhaps also be reassured that their statistical needs can still be met.

7.3 Possible Impacts on Redistricting

As indicated above, redistricting bodies will require population and other data for regions with populations infused with noise from the DP process. There is concern that the population estimates derived from differentially protected Census block data will lead to uncertainties in designing state and Congressional voting districts. Census has begun to consider these issues, for example, in their recent end-to-end test for the state of Rhode Island [40]. We cannot discuss the variance of the actual counts and those treated under DP quantitatively here as these data are protected under Title 13. But, especially for the counts associated with smaller state legislature districts, the variances may lead to concerns in verifying that the districts are properly sized relative to the requirements of the Voting Rights Act. JASON was briefed by Justin Levitt [18] that such district equalization is a "legal fiction" since it is impossible to guarantee the accuracy and precision of the counts; they are a snapshot in time and so are not temporally static. Overall, the noise from block-level estimates is not expected to lead to legal jeopardy, but could in the case where, for example, racial makeup nears thresholds that elicit concern. Census is currently engaged with the Department of Justice regarding this issue but at the time of the writing of this report, Census has not allayed the Department of Justice's concerns regarding this issue.

7.4 Limiting Release of Small Scale Data

The trade-off between probability of re-identification and statistical accuracy is reflected in the choice of the DP privacy-loss parameter. A low value increases the level of injected noise (and thus also decreases probability of re-identification) but degrades statistical calculations. Another factor that also influences the choice of privacy-loss parameter is the number and geographical resolution of the tables released, an aspect of granularity of the allowed queries. For example, if no block-level data were publicly released, a re-identification "attack" of the sort described above presumably would become more difficult, perhaps making it feasible to add less noise and so allowing a larger value of ε.

For those public officials and researchers needing access to the finer scale block level data, special channels in the form of protected enclaves may be required. We discuss this next in Section 7.5. This most likely cannot be a solution for certain uses of Census data mandated by law. For example, redistricting must be performed in a way that is transparent to the public. Today this requires using block level populations in designing the new districts. These will be infused with noise under differential privacy. While it is thought that these population estimates can still be used for redistricting, their overall utility is closely tied to the value of ε that is ultimately chosen. Too low a value of ε may lead to concern over the totals. This seems to be a particularly difficult problem that must be solved in close consultation with the relevant stakeholders.

7.5 The Need for Special Channels

Depending on the ultimate level of privacy protection that is applied for the 2020 census, some stakeholders may need access to more accurate data. A benefit of DP is that products can be generated at various levels of protection depending on the level of statistical accuracy required. The privacy-loss parameter can be viewed as a type of knob by which higher settings lead to less protection but more

accuracy. However, products publicly released with too low a level of protection will again raise the risk of re-identification.

One approach might be to use technology (e.g. virtual machines, secure computation platforms etc.) to create protected data enclaves that allow access to trusted stakeholders of census data at lower levels of privacy protection. Inappropriate disclosure of such data could still be legally enjoined via the use of binding non-disclosure agreements such as those currently in Title 13. This idea is similar to the concept of "need to know" used in environments handling classified information. In some cases there may emerge a need to communicate to various trusted parties census data either with no infused noise or perhaps less infused noise than applied for the public release of the 2020 census. Examples include the need to obtain accurate statistics associated with state or local government initiatives, or to perform socio-economic research associated with small populations.

At present, the only way to obtain data not infused with noise is to apply for access via a Federal Statistical Research Data Center. These centers are partnerships between federal statistical agencies like the Census and various research institutions. The facilities provide secure access to microdata for the purposes of statistical research. As of January 2018, there were 294 approved active projects with Census accounting for over half of these. All researchers must at present obtain Census Special Sworn Status (to uphold Title 13), pass a background check and develop a proposal in collaboration with a Census researcher.

The use of DP presents an opportunity to expand the number of people who may access more finely-grained data but who would not need to access the original microdata. Products could be constructed at higher levels of the privacy loss parameter than that used in releasing Census data to the public. In a sense, the use of DP allows Census to control the level of detail available to a researcher but in accord with the users "need to know", or more appropriately their need to access data at a given level of fidelity.

If such a program is developed there may arise the need to increase the ca-

pacity of the research data centers but at the same time the requisite security must be enforced. The defense and intelligence communities are facing similar issues and have responded by using cloud-based infrastructure and “thin client" terminals with limited input/output capability and strongly encrypted communication to ensure that data is appropriately protected and not handled improperly.

Transformative work in various areas of social science and economics has resulted from the ability to access and analyze detailed Census data. For example, Chetty and his colleagues [3] have used detailed census data to research approaches to using DP in small areas while maintaining the guarantees of DP. The development of virtual enclaves would expand opportunities to make similar contributions to a much wider cohort of researchers.

8 Conclusion

We conclude this report with a discussion of the controversy that has arisen as a result of the discovery of the Census vulnerability. The need to address the Census vulnerability also brings forward aspects of a tension between laws that protect privacy as opposed to those that require the government to report accurate statistics. We close with a set of findings and recommendations.

8.1 The Census Vulnerability Raises Real Privacy Issues

In the view of JASON, Census has convincingly demonstrated the existence of a vulnerability that census respondents can be re-identified through the process of reconstructing microdata from the decennial census tabular data and linking that data to databases containing similar information that can identify the respondent. The re-identification relied on matching Census records with commercial marketing datasets. These data providers, such as Experian, ConsumerView, and others already have a good deal of the data Census must secure such as name, age, gender, address, number in household, as well as credit histories, auto ownership, purchasing, consumer tastes, political attitudes, etc. But we note that the accuracy and granularity of their data is almost surely less than Census, and they generally do not include race or Hispanic identity; the latter is most likely a choice, not a fundamental constraint on information collection. In addition to this data there is also proprietary data maintained by Facebook, the location data collected by cell phone providers, etc.

One might argue that Census data is not of much additional utility given the limited amount of information gathered in the decennial census. However, many components of the data Census collects are not in the public domain and are still viewed as private information. For example information on children is hard to purchase commercially because its collection is enjoined by laws such as the Children's Online Privacy Protection Act. Other examples include race, number

and ages of children, sexuality of household members and, in the near future, citizenship status. Census has an obligation to protect this information under Title 13 and, in view of the demonstrated vulnerability, it is clear that the usual approaches to disclosure avoidance such as swapping, top and bottom coding, etc. are inadequate. The proposal to use Differential Privacy to protect personal data is promising although further work is requried as this report points out.

The decision to use Differential Privacy has elicited concerns from demographers and social scientists. Ruggles has argued, for example, that Census has not demonstrated that the vulnerability it discovered is as serious as claimed. In [29] he states

> "In the end only 50% of the reconstructed cases accurately matched a case from the HDF source data. In the great majority of the mismatched cases, the errors results from a discrepancy in age. Given the 50% error rate, it is not justifiable to describe the microdata as 'accurately reconstructed'."
>
> Reconstructing microdata from tabular data does not by itself allow identification of respondents allow identification of respondents; to determine who the individuals actually are, one would then have to match their characteristics to an external identified database (including, for example, names or Social Security numbers) in a conventional re-identification attack. The Census Bureau attempted to do this but only a small fraction of re-identifications actually turned out to be correct, and Abowd ... concluded that 'the risk of re-identification is small.' Therefore, the system worked as designed: because of the combination of swapping, imputation and editing, reporting error in the census, error in the identified credit agency file, and errors introduced in the microdata reconstruction, there is sufficient uncertainty in the data to make positive identification by an outsider impossible."

This statement may reflect the state of affairs prior to the re-identification ef-

fort of the Census discussed in Section 4.1 that succeeded in re-identifying 17% of the US population in 2010. An earlier re-identification attempt by the Census had some issues matching the Census geo-ids with those of commercial data. Once this was understood and fixed, the results discussed in Section 4.1 were obtained.

Ruggles also argues that use of differential privacy will mask respondents characteristics, data that are valuable in demographic and other studies. He correctly asserts that masking characteristics is not explicitly required under the law. But Census is prohibited from publishing

> "any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means..."

Given the level of re-identification that was achieved in the Census vulnerability study, it is certainly arguable that releasing tabular information without noise such that the microdata can be reconstructed and possibly matched with external data makes the tabular information just such a representation.

Ruggles further argues that Census would not validate any potential re-identification. This is true, but the fact remains that a commercial data provider can still perform the re-identification attack, then perform a probabilistic record match (perhaps using data held out from the re-identification), and, if the result looks sufficiently promising, add this to their database along with extra information on race, children, sexuality, etc. The argument that Census will not confirm the re-identification is true whether one performs any disclosure avoidance or not. But it is still the responsibility of Census not to abet such re-identification. Finally, there is the issue of whether Census data (as opposed to ACS data) is particularly sensitive. It can be argued that knowledge of various characteristics combined with location data could certainly be abused in various instances and so this provides further support that Census should enforce privacy of census data.

Even more concern has been voiced in the social science and demographer

communities regarding the possibility that the ACS tables and microdata sample may also now require similar protection. To date Census has not established that a similar vulnerability exists for the ACS data. Intuitively, it *should* be harder to re-identify this data as it is a small sample of the population and what is released is carefully chosen so as to preserve confidentiality. In any case, no plan by Census exists at present to apply methods of formal privacy to the ACS, and no changes are envisioned in the format for data release at least until 2025 when the issue will be reconsidered (cf. for example, [33]).

8.2 Two Statutory Requirements are in Tension in Title 13

It is to be expected that advances in technology may introduce tensions or conflicts among statutory provisions that were seen as conflict-free when they were enacted in the past. Under the Executive Branch's broad powers to interpret and apply the law, responsibility falls on Executive agency government officials to set policies that attempt to "square the circle" in a defensible manner, even when no perfect solution is possible. Such policies, both as to the procedure of how they are set and their substance, are potentially subject to judicial review, e.g., under the Administrative Procedures Act (5 USC Section 500). The resolution of statutory conflicts is thus ultimately a matter for the courts, or for Congress if it chooses to change the law.

In the above light, we examine two statutory provisions of Title 13. Section 214 ("Wrongful disclosure of information") provides

> "[No official] may make any publication whereby the data furnished by any particular establishment or individual under this title can be identified..."

There is little or no case law to guide us in the interpretation of what, at first sight, seems a clear provision. But how clear is it? Does "whereby" mean by itself

without reference to other sources of (e.g., commercial) data? Or does "whereby" mean may not add, even incrementally in the smallest degree, to the likelihood that an individual can be identified using commercially available data? Or is it something in-between? What about "can be identified"? Does this mean identified with certainty? Or does it mean identified probabilistically as more likely than other individuals? And, if the latter, what is the quantitative level of probability that is prohibited?

Census has traditionally adopted very strict interpretations of Section 214 for a host of good reasons, including that doing so encourages trust and participation in the census. Section 141 (Public Law PL 94-171) specifies a process by which the states propose, and the Secretary of Commerce agrees to, a geographical specification of voting districts within each state[3]. It then requires that

> "Tabulations of population for the areas identified in any plan approved by the Secretary shall be completed by him as expeditiously as possible after the decennial census date and reported to the Governor of the State involved and to the officers or public bodies having responsibility for legislative apportionment or districting of such State ... "

The plain-language meaning of "tabulation of population" is fairly obvious: one counts the number of persons satisfying some required condition(s) and enters that number into a table. At the time of the 2010 Census, and with the disclosure avoidance procedures adopted at that time, there seemed to be no significant conflict between the statutory requirements of Section 214 and Section 141. Swapping, for example, preserves population counts in any geographical area. To the extent that swapped individuals were matched for other characteristics (e.g., voting age), counts of persons with matched characteristics would also be preserved. Finally, the use of swapping may allow for the use of a larger value of ε used for

[3]Technically the law says "...the geographic areas for which specific tabulations of population are desired". This has been identified as blocks and voting districts since the law was passed

publication of the various tabulations. This would have to be determined through an empirical assessment of re-identification risk performed both with and without swapping.

Census has determined, and JASON agrees, that swapping alone is an insufficient disclosure avoidance methodology for the 2020 Census. The proposed use of DP in the 2020 Census, which is by now almost certain, will bring the mandates of Section 214 and Section 141 into conflict to a substantially greater degree than previously. Although Census proposes to impose invariants along a backbone of nested geographical regions, the revised state voting districts mayh not be on this backbone, and hence will be subject to count errors whose magnitude depends on the amount of DP imposed (i.e., the choice of ε).

There is no perfect resolution of the conflict. JASON heard the opinion of some experts outside of government that inaccuracies as large as 1000 persons in state voting district counts are acceptable. However, we also heard that, in many cases, the actions of state officials can be interpreted as indicating a mistaken belief that the counts are much more accurate than this. We are not aware of any case law or judicial guidance on the issue. Thus, Census will need to adopt a policy that is a sensible compromise between conflicting provisions of law, recognizing that the ultimate adjudication of such a policy - should it prove to be controversial - lies elsewhere. Too small a value of ε, while more perfectly satisfying Section 214, satisfies Section 141 less perfectly, both being statutory requirements.

We conclude this report with JASON's findings and recommendations.

8.3 Findings

8.3.1 The re-identification vulnerability

- The Census has demonstrated the re-identification of individuals using the published 2010 census tables.

- Approaches to disclosure avoidance such as swapping and top and bottom coding applied at the level used in the 2010 census are insufficient to prevent re-identification given the ability to perform database reconstruction and the availability of external data.

8.3.2 The use of Differential Privacy

- The proposed use by Census of Differential Privacy to prevent re-identification is promising, but there is as yet no clear picture of how much noise is required to adequately protect census respondents. The appropriate risk assessments have not been performed.

- The Census has not fully identified or prioritized the queries that will be optimized for accuracy under Differential Privacy.

- At some proposed levels of confidentiality protection, and especially for small populations, census block-level data become noisy and lose statistical utility.

- Currently, Differential Privacy implementations do not provide uncertainty estimates for census queries.

  As has been seen in Section 6, as the geographic resolution becomes finer, DP will by design affect query results. In such cases, there will at least be a need to inform users of the variances associated with a given query. While the amount of noise injected into tables is known as a result of the open publication of the privacy budgets, the variance in a query is also affected by the size of the population involved in answering that query, the use of the high-dimensional matrix method, the enforcement of invariants, etc. complicating the error analysis. Error assessment could be accomplished by performing multiple instances of a query and then assessing the variation of the results, but this requires re-accessing the data and so potentially violating the DP bounds. Ashmeade [2] has proposed an approach to

estimate query error by using the post-processed results and then assessing variance using those results. This has the advantage that one need not access the confidential data. Ashmeade presents some empirical evidence that, for the most part, this approach yields sensible bounds, but for small privacy budgets occasional outliers occur and the results of such an estimate vary widely from the true results obtained using Monte-Carlo methods. This issue clearly requires further work.

8.3.3 Stakeholder response

- Census has not adequately engaged their stakeholder communities regarding the implications of Differential Privacy for confidentiality protection and statistical utility.
- Release of block-level data aggravates the tension between confidentiality protection and data utility.
- Regarding statistical utility, because the use of Differential Privacy is new and state-of-the-art, it is not yet clear to the community of external stakeholders what the overall impact will be.

8.3.4 The pace of introduction of Differential Privacy

- The use of Differential Privacy may bring into conflict two statutory responsibilities of Census, namely reporting of voting district populations and prevention of re-identification.
- The public, and many specialized constituencies, expect from government a measured pace of change, allowing them to adjust to change without excessive dislocation.

8.4 Recommendations

8.4.1 The re-identification vulnerability

- Use substantially equivalent methodologies as employed on the 2010 census data coupled with probabilistic record linkage to assess re-identification risk as a function of the privacy-loss parameter.
- Evaluate the trade-offs between re-identification risk and data utility arising from publishing fewer tables (e.g. none at the block-level) but at larger values of the privacy-loss parameter.

8.4.2 Communication with external stakeholders

- Develop and circulate a list of frequently asked questions for the various stakeholder communities.
- Organize a set of workshops wherein users of census data can work with differentially private 2010 census data at various levels of confidentiality protection. Ensure all user communities are represented.
- Develop a set of 2010 tabulations and microdata at differing values of the privacy-loss parameter and make those available to stakeholders so that they can perform relevant queries to assess utility and also provide input into the query optimization process.
- Develop effective communication for groups of stakeholders regarding the impact of Differential Privacy on their uses for census data.
- Develop and provide to users error estimates for queries on data filtered through Differential Privacy.

8.4.3 Deployment of Differential Privacy for the 2020 census and beyond

- In addition to the use of Differential Privacy, at whatever level of confidentiality protection is ultimately chosen, apply swapping as performed for the 2010 census so that no unexpected weakness of Differential Privacy as applied can result in a 2020 census with less protection that 2010.

  There is always the possibility that unforeseen issues or implementation errors may lead to violations of the privacy protections that DP aims to enforce. Such things have happened in the past, for example, in the cryptographic community. JASON recommends that Census apply the traditional disclosure avoidance procedures applied in the 2010 census and then apply DP on top of this dataset. The advantage in JASON's view is that one can communicate that DP is a proposed improvement over traditional approaches and, should there arise any issue with DP, the previously used protections will still be in force. The software infrastructure for the traditional disclosure avoidance approach would have to be reconstructed and this could prove to be a challenge.

- Defer the choice of the privacy-loss parameter and allocation of the detailed privacy budget for the 2020 census until the re-identification risk is assessed and the impact on external users is understood.

- Develop an approach, using real or virtual data enclaves, to facilitate access by trusted users of census data with a larger privacy-loss budget than those released publicly.

- Forgo any public release of block-level data and reallocate that part of the privacy-loss budget to higher geographic levels.

- Amid increasing demands for more granular data and in the face of conflicting statutory requirements, seek clarity on legal obligations for protection of data.

A APPENDIX: Information Theory and Database Uniqueness

> Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte.
> *(I'd not have made this [letter] so long, had I had time to make it shorter.)*
>
> Blaise Pascal, *Lettres Provinciales*, 4 Dec. 1656.

In this appendix we examine the Dinur-Nissim (DN) results in the context of information theory. As a reminder, DN idealize a database as a string $d = (d_1, \ldots, d_n)$ of n bits, and a *noiseless* query as the sum of a specified subset of those bits; that is to say, the answer to the query is

$$A(q) = \sum_{i \in q} d_i \equiv \boldsymbol{w}_q^T \boldsymbol{d} \tag{A-1}$$

In the second form above, the string d is represented by a column vector $\boldsymbol{d}$, whose components are either 0 or 1, while $\boldsymbol{w}_q^T$ is a row vector of weights applied to the bits before summation; these weights are also 0 or 1, the total number of nonzero weights in $\boldsymbol{w}_q$ being denoted $\#q$, the size of the subset of bits that this query interrogates. Clearly $A(q)$ is an integer (a *count*) in the range $\{0, \ldots \#q\}$. There are of course 2^n possible distinct queries.

A.1 Noiseless Reconstruction via Linear Algebra

Each noiseless query constitutes a linear constraint on the n bits, and distinct queries obviously constitute linearly independent constraints. Here "linear" and "independent" are used in the sense of linear algebra, which therefore guarantees that n independent queries are *sufficient* to reconstruct d. Since, however, each component of d (viewed as a vector in $\mathbb{R}^n$) is restricted to only two possible values, reconstruction may be possible with fewer than n queries.

In what follows, we will often speak of the "probability" of the value of a given bit or bits in the database. In the real world, the noiseless database is fixed, so its bits are not random variables. But in order to be able to apply information-theoretic arguments to the noiseless case, let's imagine that we are designing a reconstruction algorithm to be applied to the ensemble of *all possible* databases of n bits. In this ensemble, each bit takes on the values 0 or 1 with equal frequencies ($= 1/2$). To the extent that the actual database can be regarded as having been chosen "at random," the values of its bits can be regarded as independent random variables.

With this prolog, consider a reconstruction scheme in which we first query $n/2$ disjoint pairs of bits: e.g., the k^{th} query q_k interrogates bits $2k-1$ and $2k$, for $k \in \{1,\ldots,n/2\}$. In the average over all 2^n possible data bases, since each of the two bits interrogated is ± 1,

$$A(q_k) = \begin{cases} 0 & \text{with probability } 1/4, \\ 2 & \text{with probability } 1/4, \\ 1 & \text{with probability } 1/2 \end{cases}$$

When either of the first two possibilities is realized, both bits interrogated by q_k are determined. Thus we may expect to reconstruct $n/2$ of the bits with these $n/2$ queries—a plausible result! But, we now have partial information about the remaining $n/2$ bits that belong to "ambiguous" pairs where $A(q_k) = 1$: namely, the two bits of such a pair must be distinct. There will be approximately $n/4$ ambiguous pairs. Thus a further $\sim n/4$ queries that interrogate only the first member of each such pair will resolve the remaining ambiguities. By this argument, we may reconstruct the database with no more than $\sim 3n/4$ queries. This is fewer than would suffice by the linear-algebra argument, but not by much; which suggests that the linear-algebra argument, though not rigorous, may be useful. As we show in the following subsections, however, it may be possible to do still better—i.e. fewer queries needed for noiseless reconstruction—by a logarithmic factor.

A.2 Information: An Introductory Example

To further illustrate the point, take the simple case of a 3-bit database. Let (B_1, B_2, B_3) represent these bits, $B_i \in \{0,1\}$, each with probabilities $\Pr(B_i = 0) = \Pr(B_i = 1) = \frac{1}{2}$. Consider two queries, $Q_L = B_1 + B_2$ (which interrogates the two leftmost bits) and $Q_R = B_1 + B_2$. There are of course 8 possible databases, and three possible values for each query, as shown in Table A-4 below:

B_1	B_2	B_3	Q_L	Q_R
0	0	0	0	0
0	0	1	0	1
0	1	0	1	1
0	1	1	1	2
1	0	0	1	0
1	0	1	1	1
1	1	0	2	1
1	1	1	2	2

Table A-4: Two queries on a 3-bit database

All 8 rows are equally probable. The *entropy* of the joint distribution (probability mass function or PMF) of the three bits is therefore

$$H(B_1,B_2,B_3) = -\sum_{B_1,B_2,B_3} P(B_1,B_2,B_3)\log_2 P(B_1,B_2,B_3) = -8 \times \frac{1}{8}\log_2\frac{1}{8} = 3\,,$$

as one might expect. Notice that in 6 out of 8 cases, the values of the three bits are fully determined by the values of (Q_L, Q_R). The exceptions are those in which $Q_L = Q_R = 1$, there being two bit combinations 010 and 101 that give this result. So in $3/4$ of the cases, two queries suffice to determine the bits, while in the remaining $1/4$, a third query is needed. Thus the *average* number of queries needed to reconstruct the database is[4]

$$\frac{3}{4} \times 2 + \frac{1}{4} \times 3 = 2.25 \quad \text{queries on average}$$

[4]One might ask whether it's possible to do better with a different pair of initial queries. There are 28 possibile pairs [$2^3 \times (2^3 - 1)/2$], but none does better than this pair.

Another way to look at this is to say that in $3/4$ of the cases, the two queries yield 3 bits worth of information; while in the remaining $1/4$ of the cases, the queries leave one bit's worth of ambiguity (the choice between databases 010 and 101), so that then in effect they yield only 2 bits of information. Thus the average number of bits of information yielded by these two queries is

$$\frac{3}{4} \times 3 + \frac{1}{4} \times 2 = 2.75 \quad \text{bits of information on average}$$

The joint PMF of (Q_L, Q_R), which follows from Table A-4, is

Q_L	Q_R	probability
0	0	1/8
0	1	1/8
1	0	1/8
1	1	2/8
0	2	0
2	0	0
1	2	1/8
2	1	1/8
2	2	1/8

Table A-5: Joint probability mass function of two queries.

The entropy of these two variables is therefore (combinations that have zero probability being omitted from the sum)

$$-\sum_{Q_L,Q_R} P(Q_L,Q_R)\log_2 P(Q_L,Q_R) = -6 \times \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{4}\log_2\frac{1}{4} = 2.75$$

Evidently, the entropy of the PMF of (Q_L, Q_R) coincides with the average number of bits of information gained from these two queries. This generalizes.

Looking ahead to Section A.4, the covariance of these two queries is

$$\boldsymbol{C} = \mathrm{cov}(Q_L, Q_R) = \frac{1}{4}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

and the Gaussian approximation described there predicts that

$$H(Q_L, Q_R) \approx \frac{1}{2}\log_2 \det(2\pi e C) \approx 2.88667$$

This is an overestimate (2.88667 instead of 2.75), presumably because the Gaussian approximation is not accurate for queries involving small numbers of bits. Yet it is qualitatively correct: 2 well-chosen queries on 3 bits yield > 2 but < 3 bits of information on average.

A.3 Information Gained Per Query

In the examples above, why do we do better by querying two bits at a time, and how can this be generalized?

Querying a single bit—noiselessly—reaps exactly one bit of information, because there are two possible outcomes (0 or 1), and averaged over all possible databases, these outcomes have equal frequency.

Consider now a query q that sums $\#q = m \geq 1$ bits. There are now $m+1$ possible values for the answer $A(q) = a \in \{0, \ldots, m\}$. In the data-base ensemble, the probabilities or frequencies frequencies $\{f_a\}$ of these outcomes have the binomial distribution $B(m, 1/2)$, meaning that

$$f_a = 2^{-m}\binom{m}{a}, \qquad \Rightarrow \quad \sum_a f_a = 1. \tag{A-2}$$

The formal information gained from this query is then

$$I(A) = -\sum_a f_a \log_2 f_a \tag{A-3a}$$

$$\approx \tfrac{1}{2}\log_2 m + \underbrace{\tfrac{1}{2}\log_2(\pi e/2)}_{\approx 1.047096} \equiv I_G(A) \tag{A-3b}$$

The second line is obtained by approximating the binomial distribution as a Gaussian (with mean $E(A) = m/2$ and variance $m/4$). Table A-6 shows that the Gaussian approximation is quite good even for small m—but not for $m = 0$, a point that will be important in Section A.7.

m	I	I_G
0	0	$-\infty$
1	1	1.047096
2	3/2	1.547096
16	3.04655	3.047096
128	4.547088	4.547096

Table A-6: Average information gain, in bits, from a single noiseless query that sums m bits. Second column is exact; third column is the Gaussian approximation.

What we have called $I(m)$ is also the *entropy* $H(X)$ of a binomially distributed random variable $X \sim B(m, 1/2)$. We use the notation I rather than H in this instance because we think of it as measuring the average *knowledge gained* after a query, rather than the *uncertainty* in the outcome of the query. But regardless of the interpretation, the mathematical rules governing information/entropy are the same.

A.4 Information Gained from Multiple Noiseless Queries

The preceding discussion shows that the most informative *single* query is the sum of all n bits: the information gained is $I(n) \approx 0.5 \log_2(n)$ for $n \gg 1$. But of course this is not enough to reconstruct all $n \gg \log_2 n$ bits. Clearly reconstruction requires multiple queries; but what is the minimum number? One may speculate that since a single query q that sums $\#q \sim O(n)$ bits yields $O(\log n)$ bits of information, it should follow that the minimum number of such queries required is $O(n/\log n)$. But this is not obvious, because queries are not independent unless they interrogate disjoint subsets of the n bits. Therefore their information will not simply add. In the first two schemes above, the subsets *were* independent: those queries interrogated individual bits or disjoint pairs of bits. But such "small" queries [$\#q \sim O(1)$] yield less information (at least individually) than "large" queries [$\#q \gg 1$]. And for $n \gg 1$, since we will need *at least* $O(n/\log n)$ queries to reconstruct, they cannot be entirely disjoint if they are individually

large.

Consider now two queries q_1 and q_2, and let $q_1 \cap q_2$ be the subset of bits that they have in common. If these queries are large, i.e., $\min(\#q_1, \#q_2) \gg 1$, then by the Central Limit Theorem, they are well approximated as Gaussian random variables, with means $E(q_i) = \frac{1}{2}\#q_i$ for $i \in \{1,2\}$, and covariance matrix

$$\boldsymbol{C} = \frac{1}{4}\begin{pmatrix} \#q_1 & \#(q_1 \cap q_2) \\ \#(q_1 \cap q_2) & \#q_2 \end{pmatrix}$$

(The prefactor comes from the fact that the mean-subtracted bit values are $\pm\frac{1}{2}$, whence the variance of individual bits is $\frac{1}{4}$.) It is easily seen that if the "information" of a multivariate Gaussian density function

$$P(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \frac{1}{\sqrt{\det(2\pi\boldsymbol{C})}} \exp\left(-\frac{1}{2}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{C}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}$$

is defined by $-\int P(\boldsymbol{x}) \log_2 P(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, then this information is

$$I(\boldsymbol{C}) = \log_2 \sqrt{\det(2\pi e\boldsymbol{C})}\,, \tag{A-4}$$

This reduces to the Gaussian approximation of Section A.3 for a single query, where $\boldsymbol{C} \to m/4$, a scalar. For multiple *disjoint* queries, so that $\boldsymbol{C}$ is diagonal, eq. (A-4) says that the total information is the sum of the informations gained from each query separately. If the queries are not disjoint, then at least some of the off-diagonal entries of $\boldsymbol{C}$ are positive, and none are negative, whence the determinant of $\boldsymbol{C}$ is less than the product of its diagonals: this means that the total information is less than the sum of the information obtained from the individual queries.

The goal now is to find the smallest rank r (i.e., the smallest number of queries) for which $I(\boldsymbol{C}) > n$, with the restriction that

$$\boldsymbol{C} = \frac{1}{4}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W}\,, \tag{A-5}$$

for some $n \times r$ matrix $\boldsymbol{W}$ whose entries are 0 or 1: each column of $\boldsymbol{W}$ corresponds to a query vector $\boldsymbol{w}_q$. If the information $I(\boldsymbol{C}) > n$, we can expect to be able to reconstruct "most" n-bit databases with these r queries.

Suppose, to begin with, that the entries of the matrix $\boldsymbol{W}$ are chosen at random. In this case, approximately half of the elements in each column (i.e., in each query vector) would be 1, and the remainder 0; but the excess or deficit of 1s over 0s in each column would fluctuate by $O(\sqrt{n})$. Any two distinct columns of $\boldsymbol{W}$ would have approximately $n/4$ 1s in common, so that $\sum_k W_{ik}W_{kj} \approx (n/4)(1+\delta_{ij})$. The elements of the covariance matrix would then be

$$C_{ij} = \begin{cases} n/8 + O(\sqrt{n}) & \text{if } i = j \in \{1,\ldots,r\} \\ n/16 + O(\sqrt{n}) & \text{if } i \neq j \end{cases} \tag{A-6}$$

The $O(\sqrt{n})$ are random in sign and have mean 0, so that it might be hoped that in computing $\log_2 \det \boldsymbol{C}$ for sufficiently large n, we could neglect them compared to the $O(n)$ terms. The matrix with these terms neglected is

$$\bar{\boldsymbol{C}} = \frac{n}{16}\left(\boldsymbol{I}_r + \boldsymbol{J}_r\right), \tag{A-7}$$

in which $\boldsymbol{I}_r$ is the $r \times r$ identity matrix, and the matrix $\boldsymbol{J}_r$ is entirely filled with 1s (sometimes called the "unit" matrix, although this risks confusion with the identity). Since $\boldsymbol{I}_r$ commutes with $\boldsymbol{J}_r$, the two matrices can be simultaneously diagonalized, and their eigenvalues simply add.

It is not hard to see that the eigenvectors of $\boldsymbol{J}$ have the form

$$\boldsymbol{v}_\omega = (1, \omega, \omega^2, \ldots, \omega^{r-1})^T$$

with $\omega^r = 1$, i.e. ω is any of the r^{th} roots of unity. These eigenvectors are orthogonal ($\boldsymbol{v}_\omega^\dagger \boldsymbol{v}_{\omega'} = r\delta_{\omega,\omega'}$), as is familiar from the Discrete Fourier Transform. For the trivial root $\omega = 1$, the eigenvalue of $\boldsymbol{J}$ is r, while all of the $r-1$ other roots correspond to zero eigenvalues. Therefore the eigenvalues of $\boldsymbol{I}+\boldsymbol{J}$ are

$$\{\underbrace{1,\ldots,1}_{r-1 \text{ times}}, 1+r\},$$

and it follows that

$$\begin{aligned} I(\bar{\boldsymbol{C}}) &\equiv \tfrac{1}{2}\log_2 \det(2\pi e \bar{\boldsymbol{C}}) \\ &= \tfrac{1}{2} r \log_2\left(\frac{\pi e}{8} n\right) + \tfrac{1}{2}\log_2(1+r) \\ &\approx \tfrac{1}{2} r(\log_2 n + 0.094) \qquad \text{for } r, n \gg 1. \end{aligned} \tag{A-8}$$

A.5 *m* Sequences and Hadamard Matrices

The replacement

$$\boldsymbol{C} \to \bar{\boldsymbol{C}}$$

is an approximation. But we can obtain the determinant (A-7) exactly in the special cases that $n = 2^k - 1$ through a cunning *pseudo*random choice of the query vectors: namely, *m*-sequences, a.k.a. maximum-length Linear Feedback Shift Register (LFSR) sequences [11]. In the form we need them here, they are periodic sequences of bits $b_i \in \{0,1\}$ with period $n = 2^k - 1$ and autocorrelation function

$$A(j) \equiv \sum_{i=0}^{n-1} b_i b_{i+j} = \begin{cases} (n+1)/2 & \text{when } j \equiv 0 \mod n \\ (n+1)/4 & \text{otherwise} \end{cases} \tag{A-9}$$

If we populate the columns of $\boldsymbol{W}$ with distinct circular shifts of such a sequence, then $\boldsymbol{C}$ will have almost exactly the form (A-7), the only change being that $n \to n+1$ (an even number). Then the information gained from these r queries will be exactly as in the second line of (A-8), except for the same replacement.[5]

Hadamard matrices yield similarly good correlation properties [11]. By definition, a Hadamard matrix of order n is an $n \times n$ matrix $\boldsymbol{H}$ whose entries are ± 1 and whose rows are orthogonal, so that $\boldsymbol{H}\boldsymbol{H}^T = n\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity. The order n must be 1, 2, or a multiple of 4; it is conjectured but not proved that Hadamard matrices exist for every multiple of 4. There are explicit constructions for special cases, however, and in particular for $n = p+1$ where p is a prime of the form $4k-1$ (i.e. $n \in \{4, 8, 12, 20, 24, 32, 44, 48, 60, \ldots\}$). Importantly, the first row (and first column) of the latter sort[6] of Hadamard matrix is all 1s, so it follows from the definition that each of the remaining rows has an equal number of $+1$*s* and -1*s*. It is then not hard to see that if we replace the elements H_{ij} of such a matrix with

$$W_{ij} = \frac{1}{2}(H_{\sigma(j)i} + 1)\,,$$

[5] Exact, that is, within our Gaussian approximation for the binomial query outcomes.

[6] a "cyclic" Hadamard matrix [11]

so that the j^{th} column of $\boldsymbol{W}$ is the $\sigma(j)^{\text{th}}$ *row* of $\boldsymbol{H}$ with every -1 replaced by 0, then the elements of $\boldsymbol{W}^T\boldsymbol{W}$ are

$$\sum_{i=1}^{n} W_{ij}W_{ik} = \begin{cases} n & j=k \;\&\; \sigma(j)=1, \\ n/2 & j=k \;\&\; \sigma(j)\neq 1, \\ n/2 & j\neq k \;\&\; \min(\sigma(j),\sigma(k))=1, \\ n/4 & j\neq k \;\&\; \min(\sigma(j),\sigma(k))\neq 1. \end{cases} \tag{A-10}$$

Here $\sigma()$ is any permutation of $\{1,2,\ldots,n\}$ But we do not have to use the complete permutation: we can use a part of it that selects some subset of r rows from $\boldsymbol{H}$, in which case $\boldsymbol{W}$ becomes $n\times r$, while the covariance matrix $\boldsymbol{C}\equiv\frac{1}{4}\boldsymbol{W}^T\boldsymbol{W}$ becomes $r\times r$. If this subset does not include the first row of $\boldsymbol{H}$ (the row that is all 1s), then $\boldsymbol{C}$ has exactly the form (A-7), and hence the same eigenvalues and determinant. If the first row of $\boldsymbol{H}$ is included, then the eigenvalues and determinant can be found by Cholesky decomposition $\boldsymbol{C}=\boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is lower triangular.

The diagonal entries of $\boldsymbol{L}$ are the square roots of the eigenvalues of $\boldsymbol{C}$. It turns out that when the first column of $\boldsymbol{W}$ is the first row of $\boldsymbol{H}$, the first diagonal of $\boldsymbol{L}$ is $\sqrt{n}/2$, all the rest are $\sqrt{n}/4$, and the rest of $\boldsymbol{L}$ vanishes except for the first column, in which all the elements after the first are also $\sqrt{n}/4$. In this case, all of the eigenvalues of $\boldsymbol{C}$ coincide with those of (A-7) (i.e., they are $n/16$) except for the first, which is $n/4$ in this case, but $n(r+1)/16$ in (A-7). So if $r<n$ (fewer queries than bits), it is slightly advantageous not to use the first row of $\boldsymbol{H}$, i.e. not to include the query that sums all of the bits.

A.6 The Minimal Number of Queries

We have seen that, within our Gaussian approximation at least, and neglecting $O(1)$ corrections, the information gained from $r\leq n$ noiseless queries on an n-bit database can be made as large as

$$\max(I_r)\approx\frac{r}{2}\left[\log_2 n+\log_2(\pi e/8)\right].$$

On the other hand, it follows from eq. (A-3a) that the maximum information obtained from a single query is $\max(I_1)\lesssim\log_2 n+\log_2(\pi e/2)$: we do best by sum-

ming all of the n bits. It would seem therefore that the redundancy among multiple queries can be made almost neglegible, i.e. $\max(I_r) \approx r \max(I_1)$: the information contributed by distinct queries is almost additive, apart from the different constants $\log_2(\pi e/2)$ vs. $\log_2(\pi e/8)$.

In the absence of prior constraints on the bits in the database, we must have $I_r \geq n$ in order to determine all of the bits. Thus

> *The minimum number of noiseless queries needed to reconstruct an n-bit database is at least $2n/\log_2 n$ for large n.*

We have tested this by numerical experiments with modest values of n and r, as shown in Table A-7. Using a modified hill-climbing technique, we have constructed a set of near-optimal (better than random) queries[7]. As shown in the fourth column, most of the 2^n possible databases answer our $\lceil 2n/\log_2 n \rceil$ queries uniquely, but not all. As we add queries, the number of ambiguous cases appears to drop exponentially. The third column shows the minimum number of queries needed to resolve all ambiguities. The evidence of this table suggests that the $r \sim 2n/\log_2 n$ criterion is relevant, but because exhaustion over all 2^n databases is impractical for much larger n, it is also consistent with the possibility that the minimum r/n needed to resolve all ambiguities asymptotes to a constant. This is what was found empirically in Section 4 but it's important to note that there is no guarantee that the least squares approach used there is optimal in the Shannon or information-theoretic sense.

A.7 Noisy Single Queries

Instead of the exact answer (A-1) to a query, we receive a noisy version $\hat{A}(q) = \boldsymbol{w}_q^T \boldsymbol{d} + N_q$, where N_q is a random variable independent of the database and query

[7] by attempting to maximize $\boldsymbol{W}^T\boldsymbol{W}$, with the restriction that $\boldsymbol{W}$ is $n \times r$ and its entries are all 0 or 1

n	$\lceil 2n/\log_2 n \rceil$	$r_{\min}$	uniques
8	6	6	98.4%
9	6	6	100%
10	7	7	100%
11	7	8	96.9%
12	7	9	88.7%
13	8	9	96.1%
14	8	9	94.6%
15	8	9	90.1%
16	8	10	83.5%
17	9	11	93.8%
18	9	13	88.0%
19	9	13	79.3%
20	10	14	95.8%
21	10	14	90.9%

Table A-7: Numerical experiments on noiseless queries of small databases. 2nd column is the smallest integer $\geq 2n/\log_2 n$. 3rd column is the minimum number of optimized queries needed to determine all 2^n databases uniquely. 4th is the fraction that are uniquely identified by $\lceil 2n/\log_2 n \rceil$ queries.

vectors. For convenience, the noise variables N_q and $N_{q'}$ belonging to distinct queries q and q' will be assumed independent and identically distributed.[8]

Presumably also there is a rule that a given query can be asked at most once—or if not, that the value taken by N_q is the same every time that query is asked: for if not, it would be possible to beat down the noise by asking the query repeatedly and averaging the answers.

The concept of *mutual information* $I(X,Y)$ is useful to express the knowledge that one has of a random variable X given an observation of a second variable Y, which for this application is a noisy version of X (Fig. A-1).

[8]This is not essential. In fact, the High Dimensional Matrix Method used by Census [19]) creates correlations among the N_q. As long as the noise remains independent of the database, the effect is to replace the noise covariance matrix $\sigma_N^2 \boldsymbol{I}$ in eq. (A-14) with some other (symmetric) matrix.

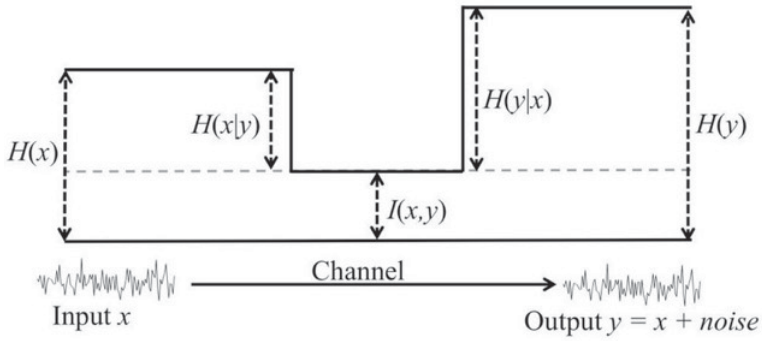


Figure A-1: Communication over a noisy channel. X ranges over transmitted signals, and Y over the noisy versions received. The entropy $H(X)$ is the minimum number of noiseless bits required to specify the value of X, and similarly for $H(Y)$. $H(X|Y)$ is the average uncertainty ($\sim$unknown bits) in X given a measurement of Y. The difference $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ is the mutual information.

The formal definition for discrete variables is

$$I(X;Y) = \sum_{X=x} \sum_{Y=y} p_{X,Y}(x,y) \log_2 \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} . \qquad \text{(A-11)}$$

Here the sums are taken over all possible values x and y of X and Y respectively, while p_X, p_Y, and $p_{X,Y}$ are the probability mass functions (PMFs) for X alone, for Y alone, and for (X,Y) jointly. It can be shown that $I(X;Y) \geq 0$, with equality iff X and Y are independent.

A small example may increase confidence in this definition. Suppose X represents a single-bit message with equally frequent values $\{0,1\}$, and $Y = X + N$ with N a noise bit that is also equally likely to be 0 or 1. Therefore $Y \in \{0,1,2\}$. The PMFs are described by the following table:

x	y	$p_X(x)$	$p_Y(y)$	$p_{X,Y}(x,y)$
0	0	1/2	1/4	1/4
0	1	1/2	1/2	1/4
0	2	1/2	1/4	0
1	0	1/2	1/4	0
1	1	1/2	1/2	1/4
1	2	1/2	1/4	1/4

The third and fourth entries in the last column (for the joint PMF) vanish, because for example if $X = 0$ then $Y = 2$ is impossible, as the noise bit is at most 1. If $Y = 0$ or $Y = 2$, then X is determined (as 0 or 1, respectively). Taken together, these outcomes happen half the time: $p_{X,Y}(0,0) + p_{X,Y}(1,2) = 1/2$. In case $Y = 1$, however, X is equally likely to be 0 or 1. So observing Y yields perfect knowledge of X half the time, and the rest of the time no information at all. We may therefore say that observing Y is worth half a bit of knowledge about X on average. If one works through the definition (A-11) using the values in this table,[9] one finds indeed that $I(X;Y) = 1/2$.

A general theorem about mutual information is[22]

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)\,,$$

in which $H(X)$ and $H(Y)$ are the entropies[10] of X and Y separately, while $H(X|Y)$ is the residual entropy of X after Y is observed, and similarly for $H(Y|X)$. This is illustrated in Fig. A-1. It is easily seen that if X and N are independent, then $H(X+N|X) = H(N)$. Therefore,

$$I(X; X+N) = H(X+N) - H(N) \quad \text{when } X \text{ is independent of } N. \tag{A-12}$$

Suppose for example that X and N are independent univariate Gaussian variables, so that $Y = X + N$ is also Gaussian, and $\mathrm{var}\,Y = \mathrm{var}\,X + \mathrm{var}\,N$. Since the

[9] It is understood that $0 \cdot \log_2 0 = 0$, i.e. cases for which $p_{X,Y}(x,y) = 0$ are excluded from the sum.

[10] See the discussion of entropy vs. information in Section A.3

entropy of a Gaussian is[11] $H(X) = \frac{1}{2}\log_2(2\pi e \operatorname{var} X)$, and similarly for $H(Y)$ and $H(N)$, it follows that

$$I(X;Y) = \tfrac{1}{2}\log_2\left(1 + \frac{\operatorname{var}(X)}{\operatorname{var}(N)}\right). \tag{A-13}$$

The logarithm here is strongly reminiscent of the factor $\log_2\left(1 + \frac{P_{\text{signal}}}{P_{\text{noise}}}\right)$ in Shannon's channel capacity theorem [31].

To relate this result to the previous discussion of noiseless queries, we need to understand what happens as the variance of the noise tends to zero. In this limit, the Gaussian approximation breaks down. The exact query results (X) are actually integers with a binomial distribution. If noise with $\operatorname{var}(N) \ll 1$ is added to such queries, the exact result (X) can be obtained from $X+N$ by rounding to the nearest integer with negligible probability of error. So we should expect $I(X, X+N)$ to reduce to $H(X)$, which is finite, as $\operatorname{var}(N) \to 0$. However, eq. (A-13) presumes that both X and N take real values, and it yields an infinite result as $\operatorname{var}(N) \to 0$ because arbitrarily close real numbers can always be distinguished.

Suppose instead that both X and N are discrete independent independent variables, for example with binomial distributions $B(m, 1/2)$ and $B(m', 1/2)$ respectively. Then $Y = X + N$ is distributed as $B(m+m', 1/2)$. Also[12] $\operatorname{var}(X) = m/4$, $\operatorname{var}(N) = m'/4$, and $\operatorname{var}(Y) = (m+m')/4$. If $m' \geq 1$, then the Gaussian approximations for $H(N)$ and $H(Y)$ are quite accurate, as shown by Table (A-6), so that eq. (A-13) is a good approximation to the mutual information. But in the noiseless case $m' = 0$, we have to use the exact definition in the first line of eq. (A-3a) for the entropy of a binomial; this yields $H(N) = 0$. Then it follows from eq. (A-12) that $I(X;Y) \to I(X;X) = H(X)$, as we expect, rather than $+\infty$ as the Gaussian approximation (A-13) would predict in the noiseless limit.

[11] For a multivariate Gaussian, this becomes $H(\boldsymbol{X}) = \frac{1}{2}\log_2 \det[2\pi e \operatorname{cov}(\boldsymbol{X})]$, where $\operatorname{cov}(\boldsymbol{X})$ is the covariance matrix of $\boldsymbol{X}$

[12] Recall that if $X \sim B(n, p)$, where p is the probability of "success" on a single trial and n is the number of trials, that $\operatorname{var}(X) = np(1-p)$.

A.8 Multiple Noisy Queries

This generalizes directly to multiple queries, represented by a vector $\boldsymbol{X}$ when exact, but corrupted by a noise vector $\boldsymbol{N}$ with diagonal covariance $\mathrm{cov}(\boldsymbol{N}) = \sigma_N^2 \boldsymbol{I}$. Provided $\sigma_N^2 \gtrsim 1/4$, we may use the Gaussian approximation, so that

$$I(\boldsymbol{X}, \boldsymbol{X}+\boldsymbol{N}) \approx \tfrac{1}{2} \log_2 \det[\sigma_N^{-2} \boldsymbol{C} + \boldsymbol{I}]. \tag{A-14}$$

in which $\boldsymbol{C} = \mathrm{cov}(\boldsymbol{X})$ is determined as before by the $n \times r$ query matrix $\boldsymbol{W}$ [eq. (A-5)], and $\boldsymbol{I}$ is the $r \times r$ identity.

The result (A-14) should be interpreted as the total information gathered by these queries in the presence of noise. As we've seen in Section A.4, for sensible (e.g. random) choices of the query matrix $\boldsymbol{W}$, all but one of the eigenvalues of $\boldsymbol{C}$ is approximately equal to $n/16$ if $n \geq r \gg 1$. It follows that the net information gathered on average is

$$I_{\mathrm{net}} \approx \frac{r-1}{2} \log_2 \left(1 + \frac{n}{16\sigma_N^2}\right) + \frac{1}{2} \log_2 \left(1 + \frac{n(r+1)}{16\sigma_N^2}\right). \tag{A-15}$$

(The second logarithm comes from the one nonzero eigenvalue of the matrix $\boldsymbol{J}$ discussed above.) If there is to be hope of reconstructing the database, the information I_{net} must be $\geq n$, the number of bits to be reconstructed. If the standard deviation of the noise $\sigma_N > \sqrt{n/48}$, however, then the logarithm < 2, in which case we will not have enough information even at $r = n$—i.e., even if we make as many queries as bits. This is reminiscent of DN's result to the effect that $O(\sqrt{n})$ noise is sufficient to prevent an "algebraically bounded" adversary from reconstructing the database.

But now suppose that we are allowed to make $r \gg n$ queries. This is most interesting in the large-noise limit, i.e. where σ_N^2 is large compared to all of the eigenvalues of $\boldsymbol{C}$. Note by the way that $\boldsymbol{C}$ becomes singular for $r > n$, because it is constructed from $\boldsymbol{W}$, which has rank $\min(r,n)$. However, the combination $\sigma_N^2 \boldsymbol{C} + \boldsymbol{I}$ is nonsingular, and for sufficiently large σ_N^2, the expansion

$$\log_e \det(\boldsymbol{I} + \varepsilon \boldsymbol{M}) \to \varepsilon \mathrm{Trace}(\boldsymbol{M}) + O(\varepsilon^2) \quad \text{as } \varepsilon \to 0 \text{ at fixed } \boldsymbol{M}$$

allows us to write

$$I_{\text{net}} \approx \frac{\log_2 e}{2\sigma_N^2}\text{Trace}(\boldsymbol{C}) \approx \frac{nr\log_2 e}{16\sigma_N^2} \quad (\sigma_N^2 \gg n/16) \tag{A-16}$$

Hence, even if the signal-to-noise ratio per query is very small, a sufficient number of queries—specifically, $r \gtrsim 16\sigma_N^2/\log_e 2$—should gather enough information to determine the database. We have not checked this prediction experimentally but we do confirm that it is possible to gather sufficient information to reconstruct the DN database provided we can issue enough queries. Note that this result indicates one will always recover the bits if the variance of the noise is held fixed as the queries are issued.

A.9 Reconstruction

So far we've talked about gathering enough information, through queries, to *determine* the bits in a database; but we haven't provided a method for actually estimating the bits from the query results. Methods based on bounded least squares optimization are discussed elsewhere in this report, and illustrated by numerical experiments. Here we provide an alternative approach, straightforwardly applying Bayesian inference to our Gaussian approximation. For simplicity, we discuss here only the noiseless case, but the method is easily generalized to include noise.

The general idea is this. We choose a full $n \times n$ matrix $\boldsymbol{W}$ of query weights, with $\det \boldsymbol{W}$ nonzero. We then ask, after the first $r < n$ of these queries (defined by the first r columns of $\boldsymbol{W}$) have been posed and answered, what is the posterior (conditional) probability distribution for the answers to the remaining $n-r$ queries that have not yet been made? If this posterior is narrow, the likely answers to the not-yet-asked queries can be predicted with probable errors less than unity (i.e., less than a bit). Then, from the results of only the first r queries, we may write down a shrewd estimate for the full $n \times n$ linear system discussed in Section A.1 and invert for the bits (rounding the real-valued answers to 0 or 1 as needed). If on the other hand the posterior is not narrow enough, we increase r (i.e., ask more

queries) until it is.

This procedure is in principle well-defined if the queries are treated exactly as discrete binomial variables. But unfortunately we do not know how to make the exact calculations except by brute force. So we resort to our Gaussian approximation. Let $\boldsymbol{X}_n$ be the full length-n vector of random variables for the outcomes of all n queries defined by some $n \times n$ weight matrix $\boldsymbol{W}_n$ with entries $\in \{0,1\}$ and $\det \boldsymbol{W}_n \neq 0$. In the Gaussian approximation, the joint distribution of $\boldsymbol{X}_n$ is determined by the means $\boldsymbol{\mu}_n = E(\boldsymbol{X}_n)$ and covariances $\boldsymbol{C}_n = E\left[\boldsymbol{X}_n - \boldsymbol{\mu}_n)(\boldsymbol{X}_n - \boldsymbol{\mu}_n)^T\right]$. As in Section A.4, since we assume uniform priors on all of the database bits (0 or 1 with equal probability), each component of $\boldsymbol{mu}_n$ equals one half the sum of the corresponding column of $\boldsymbol{W}_n$, while $\boldsymbol{C}_n = \frac{1}{4}\boldsymbol{W}_{\boldsymbol{n}}^T \boldsymbol{W}_{\boldsymbol{n}}$.

Now partition $\boldsymbol{X}_n$ into its first r components $\boldsymbol{X}_r$ and the remaining $n-r$ components $\boldsymbol{X}_{n-r}$, with corresponding partitions of the means and covariances:

$$\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_r \\ \hline \boldsymbol{\mu}_{n-r} \end{array}\right], \qquad \boldsymbol{C}_n = \left[\begin{array}{c|c} \boldsymbol{C}_r & \boldsymbol{C}_{r,n-r} \\ \hline \boldsymbol{C}_{n-r,r} & \boldsymbol{C}_{n-r} \end{array}\right] \tag{A-17}$$

Here

$$\boldsymbol{C}_r = E(\boldsymbol{X}_r \boldsymbol{X}_r^T)$$

represents the $r \times r$ covariances of the components of $\boldsymbol{X}_r$ among themselves, and similarly for

$$\boldsymbol{C}_{n-r} = E(\boldsymbol{X}_{n-r} \boldsymbol{X}_{n-r}^T);$$

while

$$\boldsymbol{C}_{r,n-r} = E(\boldsymbol{X}_r \boldsymbol{X}_{n-r}^T)$$

and its transpose

$$\boldsymbol{C}_{n-r,r} = E(\boldsymbol{X}_{n-r} \boldsymbol{X}_r^T)$$

encode the $r \times (n-r)$ cross-correlations between the components of $\boldsymbol{X}_r$ and $\boldsymbol{X}_{n-r}$. As is well known,[13] the conditional probability $\Pr(\boldsymbol{X}_{n-r} | \boldsymbol{X}_r = \boldsymbol{x}_r)$ is itself

[13] see, e.g., the Wikipedia article "Multivariate normal distribution" and references therein

Gaussian, with means and covariances

$$\hat{\mu}_{n-r} = \boldsymbol{\mu}_{n-r} + \boldsymbol{C}_{n-r,r}\boldsymbol{C}_r^{-1}(\boldsymbol{x}_r - \boldsymbol{\mu}_{n-r})$$
$$\hat{\boldsymbol{C}}_{n-r} = \boldsymbol{C}_{n-r} - \underbrace{\boldsymbol{C}_{n-r,r}\boldsymbol{C}_r^{-1}\boldsymbol{C}_{n-r,r}^T}_{\boldsymbol{Q}}. \tag{A-18}$$

Since the matrix $\boldsymbol{Q}$ is positive semidefinite, it follows that $\det\hat{\boldsymbol{C}} \le \det\boldsymbol{C}_{n-r}$, with equality only if the cross correlations $\boldsymbol{C}_{n-r,r}$ vanish.

Importantly, the reduced covariance matrix $\hat{\boldsymbol{C}}$ for the unposed $n-r$ queries does not depend on the results ($\boldsymbol{X}_r = \boldsymbol{x}_r$) of the first r queries, so we can work it out in advance in terms of the query weights $\boldsymbol{W}_n$. This can be done explicitly when $\boldsymbol{C}_n$ has the simple form (A-7), which we can obtain by choosing the columns of $\boldsymbol{W}$ to be m sequences, or by choosing them at random and neglecting the resulting $O(\sqrt{n})$ "fluctuations" in the resulting components of $\boldsymbol{C}$ [eq. (A-8)]. In this case, $\boldsymbol{C}_r$ and $\boldsymbol{C}_{n-r}$ have similar forms, except that in each case, $\boldsymbol{I}$ and $\boldsymbol{J}$ are matrices of the appropriate order.[14] It's clear that $\boldsymbol{J}_k^2 = k\boldsymbol{J}_k$ for every k, and therefore

$$(\boldsymbol{I}_k + \boldsymbol{J}_k)^{-1} = \boldsymbol{I}_k - \frac{1}{k+1}\boldsymbol{J}_k$$

The off-diagonal matrix $\boldsymbol{C}_{r,n-r} = \frac{n}{16}\boldsymbol{J}_{r,n-r}$, $\boldsymbol{J}_{k,m}$ being the $k \times m$ matrix with all entries equal to 1 (so that $\boldsymbol{J}_{k,k} = \boldsymbol{J}_k$). By means of the rules

$$\boldsymbol{J}_{j,k}\boldsymbol{I}_k = \boldsymbol{J}_{j,k} \quad \text{and} \quad \boldsymbol{J}_{i,k}\boldsymbol{J}_{k,j} = k\boldsymbol{J}_{i,j}$$

we can now evaluate the reduced covariance (A-18) for this choice of queries:

$$\hat{\boldsymbol{C}}_{n-r} = \frac{n}{16}\left(\boldsymbol{I}_{n-r} + \frac{1}{r+1}\boldsymbol{J}_{n-r}\right). \tag{A-19}$$

The determinant of $\hat{\boldsymbol{C}}_{n-r}$ is smaller than that of $\boldsymbol{C}_{n-r} = \frac{n}{16}(\boldsymbol{I}_{n-r} + \boldsymbol{J}_{n-r})$ by a factor $(2r+1)/(r+1)^2 \approx 2r^{-1}$ for $r \gg 1$. In logarithmic terms, this is a disappointingly slight reduction in uncertainty.

[14] I.e., $\boldsymbol{C}_k = \frac{n}{16}(\boldsymbol{I}_k + \boldsymbol{J}_k)$, with $\boldsymbol{I}_k$ being the $k \times k$ identity, and $\boldsymbol{J}_k$ being the $k \times k$ matrix with all elements equal to 1. The prefactor $\frac{n}{16}$ in $\boldsymbol{C}_k$, however, is invariant.

B MATLAB CODE FOR DN DATABASE RECONSTRUCTION

The MATLAB codes in this appendix can be used to generate the various figures in the report associated with the calculations on the Dinur-Nissim database.

Listing 1: Matlab script for Figure 5-1

```
% script to recover the bits in a Dinur–Nissim database without noise addition

max_n_data = 1000;
min_n_data = 1000;
step_n_data = 10;

% number of random trials

n_trials = 100;

n_entry = floor((max_n_data–min_n_data)/step_n_data)+1;

n_q_recovery = zeros(1,n_entry);
n_d = zeros(1,n_entry);
n_q_norm = zeros(1,n_entry);

completion_counter_max = 10;% the consecutive number of times the min fraction correct is 1 before terminating the queryloop

i_noise  = false; % set to false for no noise addition

i_entry = 0;

i_fig = 0;


for n_data = min_n_data:step_n_data:max_n_data

    % noise level – we add gaussian noise with mean 0 and variance eta

    sigma = sqrt(n_data)/2.0; % sigma for binomial distribution

    eta = sigma*log(n_data); % ensuring the noise is just above the sqrt(n) growth
```

```
% query_fraction = linspace(1/n_data,1.0,query_max);

% generate random data set

d = randi([0,1],n_data,1);

options = optimset('display','off'); % turn off the display

% set the lower and upper bounds on the solution

lb = zeros(n_data,1);
ub = ones(n_data,1);

fraction_correct = zeros(n_trials,10000);

i_query = 0;

completion_counter = 0;

while (completion_counter < completion_counter_max)

    i_query = i_query + 1;

    max_fraction_corrrect = 0.0;
    max_residual_norm = 0.0;

    for i_trial = 1:n_trials

        % generate the random query matrix

        Q = randi([0,1], i_query, n_data);

        % generate the query answers

        ans_q = Q*d;

        % add noise to the answers

        rand_vec = normrnd(0,eta, [i_query, 1]);

        if (i_noise)
            ans_q = ans_q + rand_vec;
        end
```

```
78
79              % now use constrained least squares to generate solution
80
81              [x_sol,res_norm,residual,exitflag,output] = lsqlin(Q,
                    ans_q,[],[],[],[],lb,ub, [], options);
82
83              max_residual_norm = max(max_residual_norm, res_norm);
84
85              % now round to 0 or 1
86
87              x_sol = round(x_sol);
88
89              % compute the percentage of bits returned correctly
90
91              n_correct = 0;
92
93              for i_bit = 1:n_data
94                  if (abs(x_sol(i_bit) - d(i_bit)) <= 1.0e-3)
95                      n_correct = n_correct +1;
96                  end
97              end
98
99              fraction_correct(i_trial, i_query)  = n_correct/n_data;
100
101         end
102
103         max_fraction_correct = max(fraction_correct(:,i_query));
104         min_fraction_correct = min(fraction_correct(:,i_query));
105
106         if ((min_fraction_correct - 0.9) >= 0)
107             completion_counter = completion_counter + 1;
108         else
109             completion_counter = 0;
110         end
111
112         fprintf (' %5i  trials  n_data: %5i query: %5i comp_counter:
                %5i  min_fraction_correct %8.4e max_frac_correct %8.4e
               max_residual: %8.4e \n', ...
113             n_trials, n_data, i_query, completion_counter,
                    min_fraction_correct, max_fraction_correct,
                    max_residual_norm)
114
115     end
116
117     n_query = i_query;
```

```
118
119     % now compute the mean percent correct and its variance
120
121     mean_fraction_correct = mean(fraction_correct);
122     var_fraction_correct = var(fraction_correct);
123
124     % now find the least value of query number that provides 100
            percent recovery
125
126     i_entry = i_entry+1;
127
128     n_d(i_entry) = n_data;
129
130     n_q_recovery(i_entry) = n_query;
131
132     for i = n_query:-1:1
133         if (abs(mean_fraction_correct(i) - 1) >= 1.0e-3)
134             n_q_recovery(i_entry) = i;
135             break;
136         end
137     end
138
139     % now produce a shaded distribution plot
140
141     x = 1:i_query;
142     y_mean = mean_fraction_correct(1:n_query);
143     y_10 = quantile(fraction_correct,0.10);
144     y_50 = quantile(fraction_correct,0.50);
145     y_90 = quantile(fraction_correct,0.90);
146
147     y_10 = y_10(1:n_query);
148     y_50 = y_50(1:n_query);
149     y_90 = y_90(1:n_query);
150
151
152     i_fig = i_fig+1;
153     figure(i_fig);
154     clf;
155
156     fprintf(' plotting figure %d...', i_fig);
157     hold on
158     plot(x,y_mean,'LineWidth',1.5);
159     plot(x,y_10);
160     plot(x,y_50);
161     plot(x,y_90);
```

```
    hold off
    title(['fraction correct vs. query for ', num2str(n_data),' bits
        with ',num2str(n_trials),' trials']);
    drawnow;
    fprintf (' plot complete\n')




end


% plot the min number of queries vs number of bits

i_fig = i_fig+1;

figure(i_fig);
clf;

plot (n_d(1:i_entry), n_q_recovery(1:i_entry));

drawnow;

% play with some possible normalizations of the min number of queries

for i_e = 1:i_entry
    n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
    %    n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
end

i_fig = i_fig+1;
figure(i_fig);
clf;

plot(n_d(1:i_entry), n_q_norm(1:i_entry));
```

Listing 2: Matlab script for Figures 5-2 and 5-3

```
% script to try to recover binary data set

max_n_data = 1000;
n_q_recovery = zeros(1,max_n_data);
n_d = zeros(1,max_n_data);
n_q_norm = zeros(1,max_n_data);

i_entry = 0;

for n_data = 100:100:max_n_data


    max_query = n_data;
    n_trials = 100;
    query_percent = linspace(1/n_data,1.0,max_query);

    % generate random data set

    d = randi([0,1],n_data,1);

    options = optimset('display','off'); % turn off the display

    % set the lower and upper bounds on the solution

    lb = zeros(n_data,1);
    ub = ones(n_data,1);

    percent_correct = zeros(n_trials,max_query);


    for i_query = 1:1:max_query

       fprintf (' n_data = %d   Performing query %d  with %d trials \
          n', n_data, i_query, n_trials)


        for i_trial = 1:n_trials

            % generate the random query matrix

            Q = randi([0,1], i_query, n_data);

            % generate the query answers

```

```
            ans_q = Q*d;

            % now use constrained least squares to generate solution

            [x_sol,res_norm,residual,exitflag,output] = lsqlin(Q,
                ans_q,[],[],[],[],lb,ub, [], options);


            % now round to 0 or 1

            x_sol = round(x_sol);

            % compute the percentage of bits returned correctly

            n_correct = 0;

            for i_bit = 1:n_data
                if (abs(x_sol(i_bit) - d(i_bit)) <= 1.0e-3)
                    n_correct = n_correct +1;
                end
            end

            percent_correct(i_trial, i_query)  = n_correct/n_data;

        end

    end

    % now compute the mean percent correct

    min_percent_correct = min(percent_correct);
    mean_percent_correct = mean(percent_correct);
    var_percent_correct = 2.0*var(percent_correct); % note I'm taking
         2 std devs
    max_percent_correct = max(percent_correct);

    % now find the lowest value of the number of queries that
        provides 100 percent recovery

    i_entry = i_entry+1;

    n_d(i_entry) = n_data;

    n_q_recovery(i_entry) = max_query;

```

```
    for i = max_query:-1:1
        if (abs(mean_percent_correct(i) - 1) >= 1.0e-3)
            break;
        else
            n_q_recovery(i_entry) = n_q_recovery(i_entry) - 1;
        end
    end

    % plot error bar plot

    figure;

    errorbar (mean_percent_correct, var_percent_correct)



end

% plot the min number of queries vs number of bits

figure;

plot (n_d(1:i_entry), n_q_recovery(1:i_entry));

% play with some possible normalizations of the min number of queries -
% here we try direct proportionality to number of bits

for i_e = 1:i_entry
    n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
end

figure;

plot(n_d(1:i_entry), n_q_norm(1:i_entry));
```

Listing 3: Matlab script for Figure 5-4

```
% script to examine the distribution of number of bits recovered for a
% fixed number of random bits in a database

max_n_data = 10;
min_n_data = 100;
step_n_data = 10;

% number of random trials

n_trials = 100;

n_entry = floor((max_n_data-min_n_data)/step_n_data)+1;

n_q_recovery = zeros(1,n_entry);
n_d = zeros(1,n_entry);
n_q_norm = zeros(1,n_entry);

completion_counter_max = 10;% the consecutive number of times the min fraction correct is 1 before terminating the queryloop

i_noise  = true; % set to false for no noise addition

i_entry = 0;

i_fig = 0;


for n_data = min_n_data:step_n_data:max_n_data

    % noise level - we add gaussian noise with mean 0 and variance eta

    sigma = sqrt(n_data)/2.0; % sigma for binomial distribution

    eta = sigma*log(n_data); % ensuring the noise is just above the sqrt(n) growth


    % generate random data set

    d = randi([0,1],n_data,1);

    options = optimset('display','off'); % turn off the display
```

```
41
42      % set the lower and upper bounds on the solution
43
44      lb = zeros(n_data,1);
45      ub = ones(n_data,1);
46
47      fraction_correct = zeros(n_trials,10000);
48
49      i_query = 0;
50
51      completion_counter = 0;
52
53      while (completion_counter < completion_counter_max)
54
55          i_query = i_query + 1;
56
57          max_fraction_corrrect = 0.0;
58          max_residual_norm = 0.0;
59
60          for i_trial = 1:n_trials
61
62              % generate the random query matrix
63
64              Q = randi([0,1], i_query, n_data);
65
66              % generate the query answers
67
68              ans_q = Q*d;
69
70              % add noise to the answers
71
72              rand_vec = normrnd(0,eta, [i_query, 1]);
73
74              if (i_noise)
75                  ans_q = ans_q + rand_vec;
76              end
77
78              % now use constrained least squares to generate solution
79
80              [x_sol,res_norm,residual,exitflag,output] = lsqlin(Q,
                    ans_q,[],[],[],[],lb,ub, [], options);
81
82              max_residual_norm = max(max_residual_norm, res_norm);
83
84              % now round to 0 or 1
```

```
85
86              x_sol = round(x_sol);
87
88              % compute the percentage of bits returned correctly
89
90              n_correct = 0;
91
92              for i_bit = 1:n_data
93                  if (abs(x_sol(i_bit) - d(i_bit)) <= 1.0e-3)
94                      n_correct = n_correct +1;
95                  end
96              end
97
98              fraction_correct(i_trial, i_query)  = n_correct/n_data;
99
100         end
101
102         max_fraction_correct = max(fraction_correct(:,i_query));
103         min_fraction_correct = min(fraction_correct(:,i_query));
104
105         if ((min_fraction_correct - 0.9) >= 0)
106             completion_counter = completion_counter + 1;
107         else
108             completion_counter = 0;
109         end
110
111         fprintf (' %5i  trials  n_data: %5i query: %5i comp_counter:
                %5i  min_fraction_correct %8.4e max_frac_correct %8.4e
               max_residual: %8.4e \n', ...
112             n_trials, n_data, i_query, completion_counter,
                    min_fraction_correct, max_fraction_correct,
                    max_residual_norm)
113
114     end
115
116     n_query = i_query;
117
118     % now compute the mean percent correct and its variance
119
120     mean_fraction_correct = mean(fraction_correct);
121     var_fraction_correct = var(fraction_correct);
122
123     % now find the least value of query number that provides 100
            percent recovery
124
```

```
    i_entry = i_entry+1;

    n_d(i_entry) = n_data;

    n_q_recovery(i_entry) = n_query;

    for i = n_query:-1:1
        if (abs(mean_fraction_correct(i) - 1) >= 1.0e-3)
            n_q_recovery(i_entry) = i;
            break;
        end
    end

    % now produce a shaded distribution plot

    x = 1:i_query;
    y_mean = mean_fraction_correct(1:n_query);
    y_10 = quantile(fraction_correct,0.10);
    y_50 = quantile(fraction_correct,0.50);
    y_90 = quantile(fraction_correct,0.90);

    y_10 = y_10(1:n_query);
    y_50 = y_50(1:n_query);
    y_90 = y_90(1:n_query);


    i_fig = i_fig+1;
    figure(i_fig);
    clf;

    fprintf(' plotting figure %d...', i_fig);
    hold on
    plot(x,y_mean,'LineWidth',1.5);
    plot(x,y_10);
    plot(x,y_50);
    plot(x,y_90);
    hold off
    title(['fraction correct vs. query for ', num2str(n_data),' bits
        with ',num2str(n_trials),' trials']);
    drawnow;
    fprintf (' plot complete\n')
```

```
end


% plot the min number of queries vs number of bits

i_fig = i_fig+1;

figure(i_fig);
clf;

plot (n_d(1:i_entry), n_q_recovery(1:i_entry));

drawnow;

% play with some possible normalizations of the min number of queries

for i_e = 1:i_entry
    n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
    %    n_q_norm(i_e) = n_q_recovery(i_e)/n_d(i_e);
end

i_fig = i_fig+1;
figure(i_fig);
clf;

plot(n_d(1:i_entry), n_q_norm(1:i_entry));
```

Listing 4: Matlab script for Figure 6-6

```
% script to examine the accuracy of a sum query as a function of the
    value
% of epsilon

n_data_row = [100 200 500 1000 2000 5000];

% number of random trials

n_trials = 1000;

trial_result = zeros(n_trials,1);

% the set of privacy loss parameters we wish to examine

eps_row = [0.001 0.005 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09
    0.1 0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19  0.2 0.3 0.4 0.5
    0.6 0.7 0.8 0.9 1.0 ];


n_d_entry = length(n_data_row);
n_e_entry = length(eps_row);


query_accuracy = zeros(n_d_entry, n_e_entry); %


for i_d_entry = 1:n_d_entry  % loop over the values of the number of
    bits

    n_data = n_data_row(i_d_entry);

    fprintf (' number of data bits: %d \n ', n_data);

    for i_e_entry = 1:n_e_entry % loop over the values of epsilon

        epsilon = eps_row(i_e_entry);

        % noise level - we add gaussian noise with mean 0 and
            variance eta

        eta =  2/epsilon^2; % this sets the variance to the
            equivalent of the two sided exponential

        for i_trial = 1:n_trials % do a number of trials to get
```

```
            reasonablre statistics
39
40          % generate random data set
41
42          d = randi([0,1],n_data,1);
43
44          % compute the correct sum
45
46          sum_query = sum(d);
47
48          % add noise to the sum of the data set - here we add a
                Laplace
49          % distribution with parameter epsilon
50
51          unif = rand() - 0.5;
52          laplace_rand_var = -1./epsilon*sign(unif)*log(1-2*abs(
                unif));
53
54  %           rand_num = normrnd(0,sqrt(eta), [1, 1]);Q_n
55
56          rand_num  =  laplace_rand_var;
57          noised_sum = round(sum_query + rand_num);
58
59          trial_result(i_trial)  = 1.0 - abs((noised_sum-sum_query)
                /sum_query); % accuray - 1 is perfect and then it
                decreases as error decreases
60
61      end
62      mean_error = mean(trial_result);
63
64      fprintf ('      epsilon = %d variance = %d mean_error=%d\n',
            epsilon, eta, mean_error);
65
66      query_accuracy(i_d_entry,i_e_entry)  = mean_error;
67
68  end
69
70  end
71
72
73
74  % now plot the results
75
76  figure;
77
```

```
78  hold on
79
80  for i_curve = 1:n_d_entry
81
82      x = eps_row;
83
84      y = query_accuracy(i_curve, 1:n_e_entry);
85
86      plot (x,y);
87
88  end
89
90  % set the axes - anything below a query accuracy of 0.0 is pretty
        useless
91  axis([0 1.0 0 1.01]);
92
93  % form the legend
94
95  for i_curve = 1:n_d_entry
96      legendCell{i_curve} = num2str(n_data_row(i_curve), 'N =%-d');
97  end
98
99  legend(legendCell);
100
101 % label the axes
102
103 xlabel('Privacy loss parameter  - \epsilon');
104 ylabel('Query accuracy');
105
106 % title the plot
107
108 title(' Dinur-Nissim query accuracy vs privacy loss parameter \
        epsilon');
```

Listing 5: Matlab script for Figure 6-7

```
1
2  % Matlab script to examine what percentage of bits are recovered for
       a given
3  % privacy loss parameter and a given number of queries in the
       presence of
4  % noise. We use a two-sided Laplace distribution to sample the noise.
5
6  % the set of database size we wish to examine
7
8  n_data_row = [4000];
9
10 % number of random trials
11
12 n_trials = 10;
13
14 trial_fraction_correct = zeros(n_trials,1);
15
16 % the set of privacy loss parameters we wish to examine
17
18 eps_row = [ 0.01 0.02 0.03 0.04 0.05 0.1 0.2 0.25 0.3 0.4 0.5 1.0  ];
19
20 % the set of multiples of the number of data points we have that we
       wish to examine
21
22 n_mult_row = [1 5 10 20];
23
24 n_d_entry = length(n_data_row);
25 n_e_entry = length(eps_row);
26 n_m_entry = length(n_mult_row);
27
28 options = optimset('display','off'); % turn off the display for the
       optimizer
29
30 % array of fraction of number of bits correct as a function of number
        of bits, number of queries, and epsilon
31 fraction_correct = zeros(n_d_entry, n_m_entry,  n_e_entry);
32
33 % loop over the values of the number of bits
34 for i_d_entry = 1:n_d_entry
35
36     n_data = n_data_row(i_d_entry);
37
38     fprintf (' number of data bits: %d \n ', n_data);
39
```

```
40      % set the lower and upper bounds on the solution
41
42      lb = zeros(n_data,1);
43      ub = ones(n_data,1);
44
45      % generate random data set
46
47      d = randi([0,1],n_data,1);
48
49      % loop over the values of epsilon
50      for i_e_entry = 1:n_e_entry
51
52          epsilon = eps_row(i_e_entry);
53
54          % noise level - we add Laplce noise with mean 0 and variance
                eta
55          % this sets the variance to the equivalent of the two sided
                exponential
56          eta =  2/epsilon^2;
57
58          fprintf ('      epsilon = %d variance = %d \n', epsilon, eta)
                ;
59
60          % loop over the queries - we do various multiples of the
                number of
61          % data points
62
63          for i_m_entry = 1:n_m_entry
64
65              i_query = n_data*n_mult_row(i_m_entry);
66
67              % we do n_trials trials and average the results
68
69              max_residual_norm = 0;
70
71              for i_trial = 1:n_trials
72
73                  % generate the random query matrix
74
75                  Q = randi([0,1], i_query, n_data);
76
77                  % generate the query answers
78
79                  ans_q = Q*d;
80
```

```
81              % add noise to the answers
82
83              % add noise to the sum of the data set - here we add
                    a Laplace
84              % distribution with parameter epsilon
85
86              unif = rand(i_query,1) - 0.5;
87              laplace_rand_var = -1./epsilon.*sign(unif).*log(1-2*
                    abs(unif));
88
89              ans_q = ans_q + laplace_rand_var;
90
91              % now use constrained least squares to generate
                    solution
92
93              [x_sol,res_norm,residual,exitflag,output] = ...
94                  lsqlin(Q,ans_q,[],[],[],[],lb,ub, [], options);
95
96              max_residual_norm = max(max_residual_norm, res_norm);
97
98              % now round to 0 or 1
99
100             x_sol = round(x_sol);
101
102             % compute the percentage of bits returned correctly
103
104             n_correct = 0;
105
106             for i_bit = 1:n_data
107                 if (abs(x_sol(i_bit) - d(i_bit)) <= 1.0e-3)
108                     n_correct = n_correct +1;
109                 end
110             end
111
112             trial_fraction_correct(i_trial)  = n_correct/n_data;
113
114         end
115
116         max_fraction_correct = max(trial_fraction_correct);
117         min_fraction_correct = min(trial_fraction_correct);
118         mean_fraction_correct = mean(trial_fraction_correct);
119         var_fraction_correct = var(trial_fraction_correct);
120
121         fprintf ('              n_data: %5i query: %5i
                mean_fraction_correct %8.4e   max_residual: %8.4e \n',
```

```
                ...
122             n_data, i_query, mean_fraction_correct,
                    max_residual_norm)
123
124         fraction_correct(i_d_entry,i_m_entry,i_e_entry)  =
                mean_fraction_correct;
125
126     end
127 end
128
129 % now plot the results
130
131 [X, Y] = meshgrid(n_mult_row, eps_row);
132
133 % loop over the size of the data vector
134
135 Z = zeros(n_e_entry, n_m_entry);
136
137 for i_d_entry = 1:n_d_entry
138
139     for i_e_entry = 1:n_e_entry
140
141         for i_m_entry = 1:n_m_entry
142
143             Z(i_e_entry, i_m_entry) = fraction_correct(i_d_entry,
                    i_m_entry, i_e_entry); % load the array of results for
                     each data set size
144         end
145
146     end
147
148     figure;
149     surf(X,Y,Z);
150     set(gca,'XScale','linear')
151     set(gca,'YScale','linear')
152
153 end
154
```

References

[1] John M Abowd. Staring Down the Database Reconstruction Theorem. Presentation to AAAS Annual Meeting Feb 16, 2019, 2019.

[2] Robert Ashmead. Estimating the Variance of Complex Differentially Private Algorithms. Presentation to Joint Statistical Meetings, American Statisticl Association, July 27, 2019, 2019.

[3] Raj Chetty and John Friedman. A practical method to reduce privacy loss when disclosing statistics based on small sample. *American Economic Review Papers and Proceedings*, 109:414–420.

[4] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. Constrained Differential Privacy for Count Data. *arXiv e-prints*, page 1710.00608, Oct 2017.

[5] Irit Dinur and Kobbi Nissim. Revealing Information while Preserving Privacy. In *PODS*, pages 202–210. ACM, 2003.

[6] Cynthia Dwork and Jing Lei. Differential Privacy and Robust Statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, pages 371–380. Association for Computing Machinery, 2009.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[8] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Found. Trends. Theor. Comput. Sci.*, 9(3-4):211–407, 2013.

[9] Ivan P. Fellegi and Alan B. Sunter. A Theory for Record Linkage. *J. Am. Stat. Assoc.*, 64(328):1183–1210, 1969.

[10] Simson Garfinkel, John M. Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data. *Commun. ACM*, 62(3):46–53, 2019.

[11] Solomon W. Golomb and Guang Gong. *Signal design for good correlation*. Cambridge, 2005.

[12] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2019.

[13] Mark Hansen. To Reduce Privacy Risk, Census Plans to Report Less Accurate Data. *New York Times*, Dec. 6, 2018.

[14] D. Kifer. Design Principles of the TopDown Algorithm. Presentation to JASON.

[15] Ios Kotsogiannis, Yuchao Tao, Ashwin Machanavajjhala, Gerome Miklau Umass, and Amherst Michael Hay. Architecting a Differentially Private SQL Engine. `http://cidrdb.org/cidr2019/papers/p125-kotsogiannis-cidr19.pdf`.

[16] Philip Leclerc. Generating Microdata with Complex Invariants under Differential Privacy. Presentation to Joint Statistical Meeting, American Statistical Association, 2019.

[17] Philip Leclerc. Results from a Consolidated Database Reconstruction and Intruder Re-identification Attack on the 2010 Decennial Census. Presentation at Workshop "Challenges and New Approaches for Protecting Privacy in Federal Statistical Programs", 2019.

[18] Justin Levitt. Uses of 2020 Redistricting Data. Presentation to JASON.

[19] Chao Li, Michael Hay, Gerome Miklau, and Yue Wang. A Data- and Workload-Aware Algorithm for Range Queries Under Differential Privacy. `http://arxiv.org/abs/1410.0265`, 2014.

[20] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *VLDB J.*, 24(6):757–781, 2015.

[21] Ashwin Machanavajjhala. Interpreting Differential Privacy. Presentation to JASON.

[22] David J. C. MacKay. *Information Theory, Inference, & Learning*. Cambridge University Press, 2003.

[23] Rachel Marks. How the 2020 Census Products Reflect Data user Feedback. Presentation to JASON.

[24] Laura Mckenna. Disclosure Avoidance for the 1970-2010 Censuses, 2018. https://www2.census.gov/ces/wp/2018/CES-WP-18-47.pdf.

[25] Ryan McKenna, Gerome Miklau, Michael Hay, and Ashwin Machanavajjhala. Optimizing Error of High-dimensional Statistical Queries Under Differential Privacy. *Proc. VLDB Endow.*, 11(10):1206–1219, June 2018.

[26] Kobbi Nissim, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, David R O'brien, and Salil Vadhan. Differential privacy: a primer for a non-technical audience. *Vanderbilt J. Entertain. Technol. Law*, page 1021596, 2018.

[27] A. Ramachandran, L. Singh, E. Porter, and F. Nagle. Exploring Re-Identification Risks in Public Domains. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*, 2012.

[28] Jerome P. Reiter. Differential Privacy and Federal Data Releases. *Annu. Rev. Stat. Its Appl.*, 6(1):85–101, 2018.

[29] Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. Differential Privacy and Census Data: Implications for Social and Economic Research. *AEA Pap. Proc.*, 109:403–408, 2019.

[30] William Sexton. Disclosure Avoidance At-Scale. Presentation to JASON.

[31] C. E. Shannon. Communication in the presence of noise. *Proc. Inst. Radio Engineers*, 37(1):10–21, 1949.

[32] Latanya Sweeney, Merce Crosas, and Michael Bar-Sinai. Sharing Sensitive Data with Confidence: the Datatags System. *Technol. Sci.*, pages 1–34, 2015.

[33] US Census Bureau. Census Bureau Continues to Boost Data Safeguards. https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html.

[34] US Census Bureau. Census End to End Disclosure Avoidance System. https://github.com/uscensusbureau/census2020-das-e2e, 2019.

[35] US Census Bureau. Census Population Density by County. https://www.census.gov/library/visualizations/2010/geo/population-density-county-2010.html, 2019.

[36] D. van Riper. Differential Privacy and the Decennial Census. Presentation to JASON.

[37] David van Riper. Differential Privacy and the Decennial Census. https://assets.ipums.org/_files/intro_to_differential_privacy_IPUMS_workshop.pdf, 2019.

[38] Victoria Velkoff. Proposed 2020 Census Data Products. Presentation to JASON.

[39] James Whitehorne. Overview of redistricting data products. Presentation to JASON.

[40] Tommy Wright. Suitability Assessment of Data treated by DA Methods for Redistricting: Observations. Presentation to JASON.

EXHIBIT 16

Laura McKenna, U.S. Census Bureau, Research & Methodology Directorate: Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Census of Population and Housing Public Use Microdata Samples (Apr. 2019)

https://www2.census.gov/adrm/CED/Papers/FY20/2019-04-McKenna-Six%20Decennial%20Censuses.pdf

Research And Methodology Directorate

Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples

By Laura McKenna

Issued April 2019

United States™ Census Bureau

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

CONTENTS

INTRODUCTION[1]

The U.S. Census Bureau conducts the decennial censuses under Title 13, U.S. Code, Section 9 mandate to not "use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports (13 U.S.C. § 9 (2007))." The Census Bureau applies disclosure avoidance (DA) techniques to its publicly released statistical products in order to protect the confidentiality of its respondents and their data.

Different DA procedures were used for the 1960, 1970, 1980, 1990, 2000, and 2010 decennial censuses' Public Use Microdata Samples (PUMS). This paper summarizes these historical methods in order to put the ongoing DA modernization effort in context. This history of decennial census disclosure avoidance methods discusses only publicly acknowledged confidentiality edits as noted in official documentation. All of the information in this summary was taken from historical public sources, except as noted. None of the information in this paper is confidential.

There is minimal public documentation of the disclosure avoidance methods used in the 1960 Census. There is no discussion of disclosure avoidance for group quarters (GQ) data in public or internal documents for the 1960, 1970, 1980, 1990, and 2000 Censuses, but the 2010 Census has an additional subsection for that purpose.[2] This paper is focused on microdata files from the censuses. The American Community Survey (ACS) is out of scope.

This history gleans procedures from various types of PUMS that differed in terms of sample size, geographic thresholds, short-form (100 percent) data vs. long-form (sample data), and universe. All publications were based on both people in households and people in GQ.

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

[2] GQ data include information about people living in nursing homes, prisons, college dormitories, and military barracks (somewhere other than a household).

MICRODATA

Statisticians use the term microdata to refer to any record-level data. At the Census Bureau, the term microdata has a narrower definition: it refers to collected data that have been cleaned, edited, and sometimes imputed so that they can be used to produce statistical tabulations and analyses. These data are presented at the record level. A microdata file consists of data at the respondent level, as opposed to aggregate counts or magnitudes. Each record represents one respondent, such as a person or household, and consists of values of characteristic variables for this respondent. Typical variables for a person-level demographic microdata file are age, race, sex, and income, and a household-level file might include mortgage payment/rent, year house built, and source of heat. Microdata files may include hundreds of such variables for each respondent.

The Census Bureau publicly releases microdata files from the decennial census and from many of its demographic and economic surveys. This paper focuses on those from previous decennial censuses. The PUMS from a decennial census is different from that of most surveys (with the exception of the ACS). The difference is due to the fact that the PUMS from the decennial census and the ACS do not contain records from each respondent. They contain records from a sample of their respondents that can be released with an underlying layer of uncertainty. The uncertainty exists from the inability to discern whether or not an individual respondent is captured in the PUMS files. This creates a scenario where a record with a unique combination of certain variables in the PUMS may not necessarily represent a unique person or household in the population (decennial census) or full sample (ACS). Microdata files from other demographic surveys contain records for all respondents.

First steps in minimizing the risk of unauthorized disclosure of microdata include removing direct identifiers such as names, addresses, and Social Security numbers. High-risk records (e.g., individuals with very large incomes or unusual jobs) are identified to ensure their visibility within the file is decreased. Other characteristics are considered for their uniqueness and their contribution to any increase in reidentification (disclosure) risk.

DISCLOSURE AVOIDANCE METHODS FOR MICRODATA

For any given microdata file, the Census Bureau has used a combination of the techniques described below.

Removing Information to Protect Microdata

Remove Direct Identifiers

Beginning with the obvious, the Census Bureau removes direct identifiers such as name, address, and telephone number.

Topcoding and Bottom-Coding

Topcoding and bottom-coding are used to eliminate outliers in a file. They are used for continuous variables such as age and dollar amounts. When topcoding, the top 0.5 percent of all values or the top 3.0 percent of all nonzero values (whichever effects the least amount of records) are cut off. They can be replaced with the topcode (cut off) value, or the mean or interpolated median of all topcoded values. At least three values must be included in the topcode or it will be lowered to meet that threshold. Bottom-codes are the same except on the other end of the distribution. An example of a bottom-coded variable might be the year that a building was built or gross income. For variables that are part of a sum, the individual parts are topcoded (or bottom-coded) prior to their summation.

Recoding and Rounding

Recoding is done for categorical and continuous variables. Each category of a variable must contain nationwide at least 10,000 weighted people or households, depending on the universe of the table. Otherwise, the category must be combined with another until the threshold rule is met. For continuous data values that the Census Bureau knows are public information (such as property taxes which has its own recoding scheme) and for some dollar amounts, recoding is also applied.

Other dollar amounts may follow one of two rounding/recoding schemes.

Round to the nearest two significant digits, or use this recoding scheme:

- Zero rounds to zero.
- 1 to 7 rounds to 4.
- 8 to 999 rounds to the nearest multiple of 10.
- 1,000 to 49,999 rounds to the nearest multiple of 100.
- 50,000 and greater rounds to the nearest multiple of 1,000.

Any totals or other derivations are calculated using the rounded numbers.

Geographic Population Thresholds

All geographic areas identified on PUMS must have a weighted population of 100,000 or more. When figuring out the population of an identified area, all geography-related variables on the file must be cross-tabulated to obtain the final population count. For example, other geographic variables may be urban/rural, Metropolitan Statistical Area (MSA) status, and other geographic areas named such as Congressional District. All geographic pieces identified after crossing all geographic variables must meet the required threshold for that PUMS.

Altering Information to Protect Microdata

Data swapping, the generation of partially synthetic data, and noise infusion are current methods for the protection of frequency count data from the decennial census and ACS. While the three methods are used mainly to protect tables for very small geographic areas, they are performed on the underlying microdata before tabulation. The PUMS files are sampled from the altered data.

Data Swapping for Household Data

The purpose of any swapping methodology is to introduce uncertainty into the data so that the data user doesn't know whether real data values correspond to certain respondents. Household records with a high risk of disclosure are typically identified through software and called uniques because they have a unique combination of certain variables. The unique records are targeted for data swapping. In the swapping procedure, a small percentage of records are matched with other records in the same file on a set of predetermined variables used as swapping attributes. A set of other variables are then swapped between the two records without disturbing the responses for nonsensitive and nonidentifying fields. The variables may be continuous or categorical. A household record is typically swapped with another household within a large area but in a different smaller area within the larger one, for example, across tracts but within the same county. Again, the swapping technique is targeted to protect frequency count tables from censuses and ACS, but the PUMS files are sampled from the swapped data, and this adds a small amount of confidentiality protection (Zayatz, 2002 and 2003) to the microdata.

Partially Synthetic Data

Applying data swapping to GQ data does not work well. Imagine swapping a nursing home (or someone who lives there) with a college dorm (or someone who lives there). The resulting data would make no sense, so the Census Bureau relies on the generation of partially synthetic data to protect GQ data from the decennial census and ACS.

The original data are modeled using a general linearized model. The process then continues with identifying unique records by cross-tabulating certain values and flagging records in the resulting cells with a count of one. Because these are GQ data, the uniques represent people rather than households. Those variable values that are causing the disclosure risk problem in a given unique record are then blanked and replaced with values generated from the model. Geography and type of GQ are never altered, and the numbers of people of less than age 18 and age 18 or more are never changed. Occasionally, a modeled (simulated) value may coincidentally be the same as the original value. Again, the partially synthetic data generation technique is targeted to protect frequency count tables from censuses and ACS, but the PUMS files are sampled from the partially synthetic data, and this adds a small amount of protection to the microdata.

Noise Infusion

At this time, noise infusion is not widely used for the protection of microdata. It is used to hide very unusual characteristics of a person or household at a given point in time that is not caught by the 10,000 threshold rule for individual categories described above. For example, consider a person who gave birth to seven children at one time, or a person who is a practicing physician at the age of 15 (both very unusual circumstances that would probably be in the news). Also very large households may present a disclosure risk. Editing procedures capture and alter many, but not all, of these unusual occurrences. The

Census Bureau does not publicly describe precisely how noise is added to protect this type of data.

1960

PUMS Data

Decennial censuses gather information from questions asked of the entire population, or from those same questions, as well as many others, asked of only a sample of the population. Those questions asked about every person and household are called 100 percent (or short-form) questions. The other questions are called sample (or long-form) questions. In the 1960 Census, 1 in 4 households received the long-form questionnaire.

The Census Bureau was the first statistical agency to publicly release microdata files (Ruggles, 2013). In 1962, the Census Bureau drew a sample of the long-form data records that would represent 1.0 percent of the population nationwide. The Census Bureau published two microdata files in the form of punch cards from those records, both using a geographic population threshold of 250,000 for each area identified. Areas that were identified on the PUMS files were called Public Use Microdata Areas (PUMAs). Areas could not cross state lines. The first file contained records from a 1-in-1000 sample of the population, and the second contained a 1-in-10,000 sample of the population. The second file was a subset of the first file. The smaller file was published for data users who may not have had the computer power or the need to process the larger file. In 1973, DualLabs published the records from the full 1 percent sample, which was recoded to match the record layout and categories of the 1970 public-use samples, <http://users.hist.umn.edu/~ruggles/JSM2005-000189.pdf>. The files contained person-level and household-level information, with persons linked to their households. Demographic data users were immensely pleased to have these files because they gave researchers the ability to retabulate and manipulate data without constraints imposed by a fixed set of predefined, printed tables, <www.icpsr.umich.edu/icpsrweb/icpsr/series/13>.

DA Techniques

The only DA techniques used for these files were the removal of direct identifiers and a geographic population threshold of 250,000.

1970

PUMS Data

In 1970, there were two long forms with some overlapping questions and some different questions. One long-form questionnaire was sent to 15 percent of U.S. households and GQ individuals, and the second was sent to 5 percent of households and GQ individuals. Six PUMS files were released from the 1970 Census. See Appendix A for an illustration. For both the 15 percent and the 5 percent long-form data, three PUMS files were released for different types of geographic areas: a file based on areas within a state, a file based on county groups mainly determined by MSAs, which can cross state lines, and a file based on neighborhood characteristics, <https://usa.ipums.org/usa/resources/codebooks/1970_pums_codebook.pdf>.

All six PUMS files were based on stratified samples of the two long-form datasets and were nonoverlapping in terms of households and people. They each contained data on 1 percent of the population nationwide. They were self-weighting. Each person or household had a weight of 100. For all six PUMS files, subsamples were drawn that represented 0.1 percent and 0.01 percent of the population to accommodate users who could only handle smaller files.

DA Techniques

All direct identifiers were removed from the files. A geographic population threshold of 250,000 per identified area was imposed and for the neighborhood characteristics files, the only geographic areas directly identified were census region and census division due to the fact that neighborhood characteristics can divide geographic areas into smaller pieces.

1980

PUMS Data

In 1980, there was just one long form that was sent to approximately 1 in 5 households. Individuals living in GQ and vacant units were also sampled. PUMS files included a 5 percent, state-based file (Sample A); a 1 percent, MSA-based file (Sample B); and a 1 percent, state-by-urban/rural-based file (Sample C). See Appendix B for a summarization of these samples and a comparison between the 1970 and 1980 PUMS. All three 1980 PUMS files were based on stratified samples of the long-form dataset, and were nonoverlapping in terms of households and people. They were self-weighting. Each person or household in the 1 percent files had a weight of 100, and each person in the 5 percent file had a weight of 20. The files had the same subject content. For users desiring smaller files, a subsample of 0.1 percent of

the population was also released for each of the three files, <https://www2.census.gov/prod2/decennial/documents/D1-D80-PUMS-14-tech.pdf/>.

All missing data values were allocated (imputed) and allocation flags for each variable were included in the PUMS. Washington, DC, and Puerto Rico were treated as states.

DA Techniques

All direct identifiers were removed from the files. Income was grouped into $10 intervals and was topcoded at $75,000, and age was topcoded at 90.

A geographic population threshold of 100,000 per identified area (PUMA) was imposed. PUMAs were not always contiguous. PUMAs in the state-based file (Sample A) could not cross state boundaries. Many PUMAs in the MSA-based file (Sample B) did cross state boundaries. PUMAs in the state by urban-rural file (Sample C) sometimes had to combine states. The PUMAs for this file consisted of the cross tabulation of state by urban/rural designation. If there was not enough population designated as urban or rural in a given state to meet the 100,000 population threshold for a PUMA, that state was combined with another. Region and division boundaries were not crossed.

1990

PUMS Data

In 1990, there was just one long form that was sent to approximately 16 percent of all U.S. households. Individuals living in GQ and vacant units were also sampled. People sampled from within the same GQ were not identifiable as such. PUMS files included a 5 percent, state-based file (Sample A); a 1 percent, MSA-based file (Sample B); and a 3 percent, elderly file for households with at least one person of age 60 or more (Sample C). All three PUMS files were based on stratified samples of the long-form dataset and were nonoverlapping in terms of households and people. Each household and person record was assigned its own weight. The files had the same subject content. For users desiring smaller files, a subsample of 0.1 percent of the population was also released for each of the three files, <https://www2.census.gov/prod2/decennial/documents/D1-D90-PUMS-14-techm.pdf>. Washington, DC, and Puerto Rico were treated as states.

In 1990, three different household sampling rates were used for the long form: 1 in 8, 1 in 6, and 1 in 2 for an overall average of about 1 in 6. For GQ, there was only one sampling rate of 1 in 6, <www.census.gov/history/pdf/1990proceduralhistory.pdf>. The variable sampling rates were used to arrive at high-quality estimates for tables published for small geographic areas and to decrease respondent burden for densely populated areas. The rates were based on precensus population estimates of incorporated places, census tracts, and block-numbering areas.

All missing data values were allocated (imputed), and allocation flags for each value were included in the PUMS. “In rare instances during the implementation of the sample weighting process, the sample size was considered inadequate to make estimates of sample data. In collection block groups with a designated 1-in-2 sampling rate, augmentation was employed if the 100 percent housing unit count was at least six and the observed sampling rate was less than 1 in 12. In collection block groups with a designated 1-in-6 or 1-in-8 sampling rate, augmentation was employed if the 100 percent, housing unit count was at least 12 and the observed sampling rate was less than 1-in-30. Augmentation was performed separately for GQ persons using the same criteria as for the 1-in-6 or 1-in-8 designated sampling rates. Augmentation was achieved by selecting a sample of housing units (or GQ persons) to increase the observed sampling rates to at least 1 in 12 or 1 in 30. Using the 100 percent characteristics, the selected households (or GQ persons) were matched by a hot deck procedure to similar housing units (or GQ persons) with sample data. The sample data were then copied to the augmented housing units (or GQ persons). The augmentation rate was very small. Most augmentation occurred for GQ persons,” <https://www2.census.gov/prod2/decennial/documents/D1-D90-PUMS-14-techm.pdf>. Augmentation is sometimes referred to as whole household imputation or, for GQ, whole person imputation.

DA Techniques

All direct identifiers were removed from the files.

The Census Bureau limited the detail on files by using recodes and topcodes for place of residence, place of work, type of GQ, income values, age, and other selected items to further protect the confidentiality of the data. Most economic items were topcoded on a national basis. The topcode was set at either 0.5 percent of all values or 3 percent of all nonzero values, whichever was the larger of the two cutoff values. If a state had at least 30 cases above a given topcode, the state median of all topcoded values was released.

A geographic population threshold of 100,000 per identified area (PUMA) was imposed for all three samples. PUMAs were not always contiguous. PUMAs in the state-based file (Sample A) did not cross state boundaries. PUMAs in the MSA-based file (Sample B) often did cross state boundaries. The elderly file (Sample C) was produced for states only. Region and division boundaries were not crossed by any PUMAs.

A confidentiality edit was performed on the underlying 1990 sample data prior to publication of all data products. It was mainly used to protect data in tables that were published for very small geographic areas, but it also affected the PUMS files. An imputation methodology was used to provide DA for sample data in small block groups. This methodology involved the blanking of a sample of the data values (population and housing items) for one of the sample housing units in each small block group and imputing those values using the 1990 Census imputation methodology. This technique was known as Blank and Impute. Once sample data imputation was completed, the resulting sample data file (for which DA had been applied) was used to prepare all subsequent census sample data products. This data imputation methodology for providing DA for sample data added very little to the level of error of the estimates. A major reason for this is that the relative increase in imputation rates was very small (Griffin et al., 1989).

2000

PUMS Data

In 2000, there was just one long form that was sent to approximately 16 percent of all U.S. households. Individuals living in GQ and vacant units were also sampled. People sampled from within the same GQ were not identifiable as such. Households and GQ people in outlying areas, such as Guam and the U.S. Virgins Islands, all received the long form. The PUMS files included a 5 percent and a 1 percent file, both state-based. Both PUMS files were based on stratified samples of the long-form dataset, and were nonoverlapping in terms of households and people. Each household and person record was assigned its own weight, <www.census.gov/prod/cen2000 /doc/pums.pdf>. Washington, DC, and Puerto Rico were treated as states. The 5 percent file identified PUMAs with a geographic population threshold of 100,000. The 1 percent file identified Super-PUMAs with a geographic population threshold of 400,000. The 1 percent file had much more variable detail (less recoding) than the 5 percent file, hence the larger areas. PUMAs and Super-PUMAs were not always contiguous, and they did not cross state boundaries. There were also PUMS files for Guam and the U.S. Virgin Islands that contained records from 10 percent of the population and had the same level of detail as the 5 percent file.

As in 1990, three different household sampling rates were used for the long form: 1-in-8, 1-in-6, and 1-in-2 for an overall average of about 1-in-6. For GQ, there was only one sampling rate of 1 in 6. The variable sampling rates were used to arrive at high-quality estimates for tables published for small geographic areas and to decrease respondent burden for densely populated areas. The rates were based on precensus population estimates of incorporated places, census tracts, and block-numbering areas.

All missing data values were allocated (imputed), and allocation flags for each value were included in the PUMS. Also as in 1990, “In rare instances during the implementation of the sample weighting process, the sample size was considered inadequate to make estimates of sample data. In collection block groups with a designated 1-in-2 sampling rate, augmentation was employed if the 100 percent, housing unit count was at least six and the observed sampling rate was less than 1 in 12. In collection block groups with a designated 1-in-6 or 1-in-8 sampling rate, augmentation was employed if the 100 percent, housing unit count was at least 12 and the observed sampling rate was less than 1-in-30. Augmentation was performed separately for GQ persons using the same criteria as for the 1-in-6 or 1-in-8 designated sampling rates. Augmentation was achieved by selecting a sample of housing units (or GQ persons) to increase the observed sampling rates to at least 1 in 12 or 1 in 30. Using the 100 percent characteristics, the selected households (or GQ persons) were matched by a hot deck procedure to similar housing units (or GQ persons) with sample data. The sample data
were then copied to the augmented housing units (or GQ persons). The augmentation rate was very small. Most augmentation occurred for GQ persons,” <https://www2.census.gov/prod2/decennial /documents/D1-D90-PUMS-14-techm.pdf>. Augmentation is sometimes referred to as whole household imputation or, for GQ, whole person imputation.

DA Techniques

All direct identifiers were removed from the files.

The Census Bureau limited the detail on files by using recodes, topcodes, and bottom codes for place of

residence, place of work, type of GQ, income values, age, and other selected items to further protect the confidentiality of the data. Most economic items were topcoded on a national basis. The topcode was set at either 0.5 percent of all values or 3.0 percent of all nonzero values, whichever was the larger of the two cutoff values. The topcode had to include at least three values in each state. If not, the topcode for a given state was lowered until the threshold was met. The mean of the topcoded values for each state was shown on the files.

Data swapping was performed on the underlying 2000 sample data prior to publication of all data products. It was mainly used to protect data in tables that were published for very small geographic areas, but it also affected the PUMS files. Once data swapping was completed, the resulting sample data file was used to prepare all subsequent census sample data products.

All categories of categorical variables on the file had to represent a nationwide universe of 10,000 weighted people. Another technique used to protect the data was noise infusion for large households. These techniques are described in detail in the section on Disclosure Avoidance Methods for Microdata.

2010

PUMS Data

In 2010, the long form had been replaced with the ACS. A PUMS file was still released from the 100 percent (short-form) data. The file was state-based and contained records for a systematic sample of 10 percent of the population in each state, Washington, DC, and Puerto Rico. Individuals living in GQ and vacant units were included. Persons sampled from within the same GQ were not identifiable as such. All missing data values were allocated (imputed), and allocation flags for each value were included in the PUMS. Each housing unit and person record was assigned a weight, <https://www2.census.gov/census_2010/12-stateside_pums/0tech_doc/2010%20pums%20technical%20documentation.pdf>. Whole household imputation and whole person imputation (for GQ) were performed.

DA Techniques

All direct identifiers were removed from the PUMS. All PUMAs had a geographic population threshold of 100,000. All categories of categorical variables on the file had to represent a nationwide universe of 10,000 weighted people. Other techniques used to protect the data included data swapping for household data, partially synthetic data generation for GQ data, topcoding, bottom-coding, recoding, and noise infusion for large households. These techniques are described in detail in the section on Disclosure Avoidance Methods for Microdata.

CONCLUSION

The content of PUMS and DA techniques have evolved over the last six censuses. All of the PUMS files contained data on individuals living in both households and GQ. Direct identifiers were removed from all of the files, and they all had geographic population thresholds. Sample (long-form) data were used to create the 1960 through 2000 files, and 100 percent (short-form) data were used to create the 2010 file. Other DA techniques for the PUMS are summarized in the table on page 10.

Recently, the Census Bureau has embarked on an aggressive effort to replace its legacy DA methods with modern DA techniques based on formal privacy methods, <https://privacytools.seas.harvard.edu/formal-privacy-models-and-title-13>. Current methods will gradually change with the introduction of formal privacy (Nissim et al., 2018). Most of the current Census Bureau's DA research is focused on formal privacy for all types of data (Nissim et al., 2007). An algorithm operating on a private database of records satisfies formal privacy if its outputs are insensitive to the presence or absence of any single record in the input (Dwork, 2006). Census Bureau staff members are quickly learning about formal privacy and how it protects Census Bureau data products.

REFERENCES

C. Dwork, "Differential Privacy," International Colloquium on Automata, Languages, and Programming (ICALP), 2006, pp. 1–12.

R. Griffin, F. Navarro, and L. Flores-Baez, "Disclosure Avoidance for the 1990 Census," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1989, pp. 516–521.

K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing, 2007, pp. 75–84.

K. Nissim, T. Steinke, A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, D. O'Brien, and S. Vadhan, "Differential Privacy: A Primer for a Non-technical Audience (Preliminary Version), Harvard University Privacy Tools for Sharing Research Data, 2018, <http://privacytools.seas.harvard.edu>.

Decennial censuses	Topcodes and Recodes	Blank and impute	Swapping	Category size thresholds	Noise infusion	Partially synthetic data
1960						
1970						
1980	X					
1990	X	X				
2000	X		X	X	X	
2010						
Households	X		X	X	X	
Group quarters	X			X		X

S. Ruggles, "Big Microdata for Population Research," Minnesota Population Center, University of Minnesota, Working Paper No. 2013-04, 2013.

L. Zayatz, "SDC in the 2000 U.S. Decennial Census," In: Domingo-Ferrer, J. (eds) Inference Control in Statistical Databases, From Theory to Practice (Lecture Notes in Computer Science), Springer, Berlin, Heidelberg, 2002, vol. 2316.

L. Zayatz, "Disclosure Limitation for Census 2000 Tabular Data," Working Paper #15, Joint ECE/Eurostat work session on statistical data confidentiality, 2003, <www.unece.org/stats/documents/2003/04 /confidentiality/wp.15.e.pdf>.

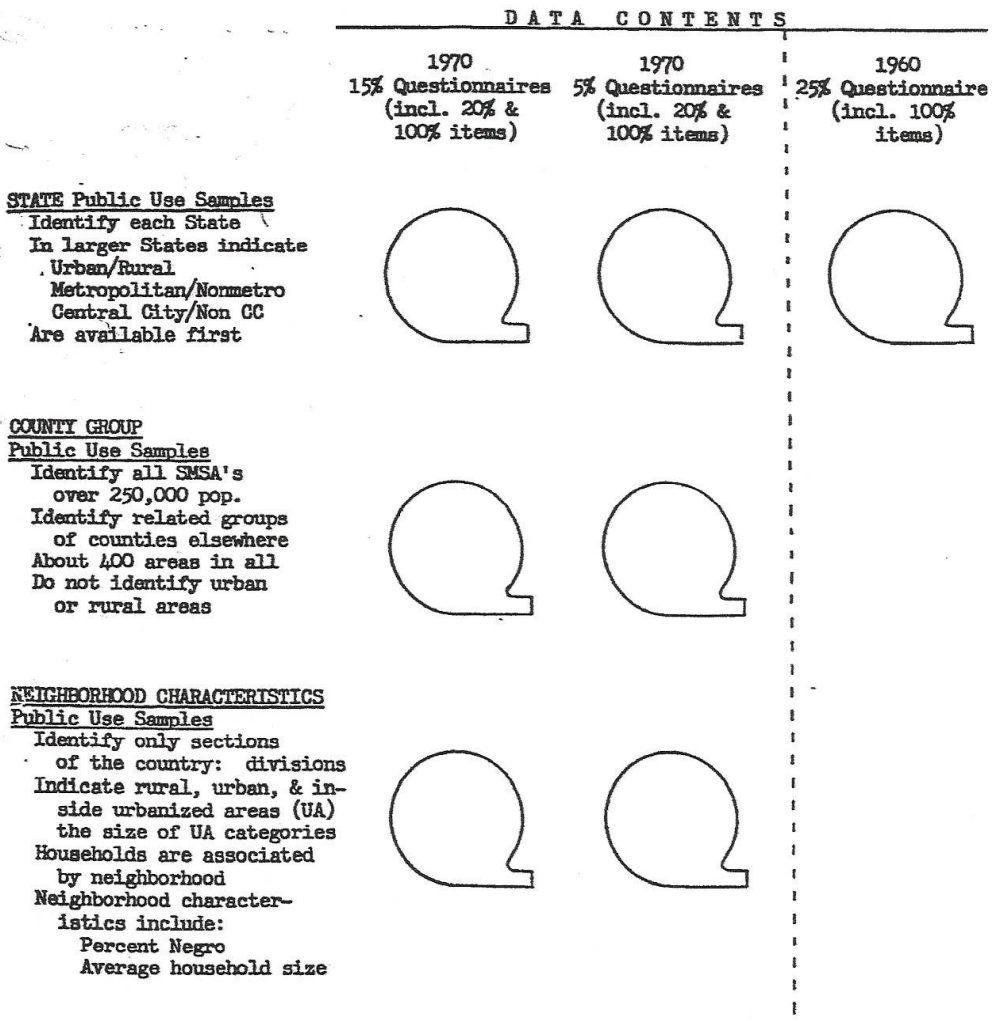
Appendix A

Public Use Samples of Basic Records from the 1970 Census

Description and Technical Documentation, p. 3

Prepared by the U.S. Census Bureau, 1972

Figure 1. PUBLIC USE SAMPLES FROM 1960 AND 1970 CENSUSES



Each tape symbol represents a separate one-in-a-hundred public use sample (30-33 tapes) from which one-in-a-thousand and one-in-ten-thousand subsamples (3 tapes and one tape respectively) will also be available.

Appendix B

1980 Census of Population and Housing

Public Use Microdata Sample

Technical Documentation, pp. 1 and 6

Prepared by the U.S. Census Bureau, 1983

CHAPTER 1. INTRODUCTION

Overview

Public-use microdata samples are computer tapes which contain records for a sample of housing units, with information on the characteristics of each unit and the people in it. In order to protect the confidentiality of respondents, the Bureau excludes identifying information from the records. Within the limits of the sample size and geographic detail provided, these tapes permit users with special needs to prepare virtually any tabulations of the data they may desire.

Three separate public-use microdata samples are available, each representing five percent or one percent of the population and housing of the United States:

- o A Sample, 5%, identifying all States and various subdivisions within them, including most counties with 100,000 or more inhabitants;
- o B Sample, 1%, identifying all metropolitan territory and most SMSAs individually, and groups of counties elsewhere;
- o C Sample, 1%, identifying regions, divisions, and most States by type of area (urban/rural).

Three 1-in-1,000 samples are also prepared, one each extracted from the A, B, and C Samples.

Figure 2. Comparison of Features on 1980 and 1970 Public-Use Microdata Samples

	1980 Samples			1970 Samples		
	A	B	C	State	County Group	Neigh Chars
Sample Size	5% 0.1%	1% 0.1%	1% 0.1%	1-2% 0.1%	1-2% 0.1%	1-2% 0.1%
Areas Identified						
Divisions	X	-	X	X	-	X
States	51	20	28	51	4	-
SMSAs of 100,000+	180	282	-	-	125	-
Counties of 100,000+	350	236	-	-	104	
Places of 100,000+	123	135	58	-	12	5
County Groups	1154	1258	-	-	409	
Urbanized Areas	-	-	73	-	-	6
Metro/Nonmetro	-	X	-	23 States	-	-
Urban/Rural	-	-	X	42 States	-	X
Neighborhood Characteristics	-	-	-	-	-	X

EXHIBIT 17

To: Earl N Mayfield (CENSUS/DEPDIR FED)[earl.n.mayfield@census.gov]; Steven K Smith (CENSUS/DEPDIR FED)[steven.k.smith@census.gov]; Benjamin A Overholt (CENSUS/DEPDIR FED)[benjamin.a.overholt@census.gov]
From: Nathaniel Cogley (CENSUS/DEPDIR FED)[/o=ExchangeLabs/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=d79dc9ef4b634b25b2efa42ed4febd8a-Cogley, Nat]
Sent: Thur 10/22/2020 6:34:51 PM (UTC)
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.

I've added Dr. Overholt to the discussion, in case he would like to add some thoughts.

—Nathaniel

From: Nathaniel Cogley (CENSUS/DEPDIR FED) <nathaniel.cogley@census.gov>
Sent: Thursday, October 22, 2020 2:28 PM
To: Earl N Mayfield (CENSUS/DEPDIR FED) <earl.n.mayfield@census.gov>; 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.

My thoughts concern these two sentences:

"We are making changes to the 2020 DAS geographic hierarchy to more effectively ensure accuracy for AIAN tribal areas. We are also working to improve the accuracy of population counts in AIAN tribal areas by allocating more of the privacy-loss budget to those statistics. "

I am not sure that any final decisions have been made. I am also not sure the logic of committing to more accurate counts for one demographic group (or type of jurisdiction) over others. It seems like the principles in question should apply to all Americans regardless of their demographic group or jurisdiction.

As people know, I am not in favor of intentionally distorting population counts anywhere and at any level of geography. I believe we have a legal obligation to accurately produce the count as determined by the decennial census.

—Nathaniel

Nathaniel Cogley, Ph.D.
Deputy Director for Policy
U.S. Census Bureau

From: Earl N Mayfield (CENSUS/DEPDIR FED) <earl.n.mayfield@census.gov>
Sent: Thursday, October 22, 2020 1:50 PM
To: Steven K Smith (CENSUS/DEPDIR FED) <steven.k.smith@census.gov>; Nathaniel Cogley (CENSUS/DEPDIR FED) <nathaniel.cogley@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.

I defer to Nathaniel--are we properly describing the data levels to which DA will apply?

Trey Mayfield
Counsel to the Director of the United States Census
Office (301) 763-0707
Cell (202) 868-1157

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.

Are we still holding on this one? Did you have further edits?

I edited it again. The tribal consultation is now scheduled.

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 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.

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>
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.j.stanley@census.gov>
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

EXHIBIT 18

2020 Disclosure Avoidance System (DAS)

Executive Guidance Group (EGG) Status Report

October 7, 2020

2020 DAS: Report Summary

Status	Report Title	Summary	Slide Number
	Group I Data Products	Development of the Disclosure Avoidance System (DAS) for the Group IA and 1B data products	4 – 15
	Group II Data Products	Development of the DAS for the Group II data products	16 – 20
	Communications and Outreach	Status how we are communicating with external stakeholders	21 – 22
	Data Products Schedule	Overview of planned public release of the 2020 census data products	23
	Budget	Budget established to support required work through fiscal year 2020	24
	Project Risks	Identification of project challenges that management continues to resolve	25
	Scope	Data products (Groups I, II, and III) to be produced using the Differential Privacy disclosure avoidance method for the 2020 census; Schedule planning information	26 – 30

Legend	Not Applicable	Completed	On-Track	Management Focus	Requires Attention

2020 DAS

Bottom Line Up Front (BLUF)

Discussion Topics

DAS Schedule Updates: Re-issue of PPMF (vintage 20200917)

Recommendations for accuracy/privacy for PL 94-171 from Census Redistricting & Voting Rights Data Office (CRVRDO)

Informational/Situational Awareness

Establish executive priorities for Sprint leading up to Operation Readiness Review (ORR)

- Prepare deliverables and checklists for Production Readiness Reviews (PRR)
- Anticipate revised ORR date in November (pending Decennial schedule CR)
- Plan for potential Disclosure Avoidance implementation for Presidential Memorandum

Develop initial timeline for Group II Products for second contract year

Release publicly Detailed Summary Metrics and Privacy-Protected Microdata File (PPMF)

Group I Data Products

2020 DAS

Business Executive Priorities

2020 DAS Sprint VIII (Sept 30 – Oct 28). ORR – TBD Nov 2020

1. **Improve privacy-accuracy tradeoff in TopDown Algorithm (TDA)**
   a) Ensuring DHC quality with PL94-171 run first
   b) Finish testing & extending Rounder improvements
   c) OLS-improved multipass
   d) Produce PPMF prior to DSEP Privacy-Loss Budget (PLB) decision (inter-sprint activity)
2. **Refine geographic spine optimization approach for optimizing PLB to better target Minor Civil Divisions (MCDs)/Census Place**
3. **Plan for potential implementation of Disclosure Avoidance for unauthorized immigrants for apportionment (Presidential Memorandum)**
4. **Prepare for Operational Readiness Review (ORR)**
   a) Create MDF20_PER.txt, MDF20_UNIT.txt, MDF20_HASHES.txt, Noisy Measurements Files (NMF20_PERssccc.txt, NMF20_UNITssccc.txt, NMF20_HASHES.txt)
   b) Finalize implementation of random number generator for differential privacy
   c) Implement data vintage system
   d) Automated restarts
   e) Test error detection and various fail safe mechanisms for optimizer failing and simple CEF errors
5. **Application Engineering Priorities**
   a) Storage and archival of noisy measurement files (NMF)
   b) Address technical debt
   c) Capture machine performance statistics for Gurobi
   d) S3 Clean-up
   e) Evaluate use of AWS EMRFS (EMR File System)

2020 DAS: Group I Status

Accomplishments to Date

Completed Sprint VII

- Modified the schema to PL94-171 only
- Implemented NNLS and Rounder improvements to TDA
- Implemented improvements to the geographic spine

Determined Disclosure Avoidance method to apply for unauthorized immigrants for apportionment (Presidential Memorandum)

Publicly released PPMF, Detailed Summary Metrics, and associated newsletter

- Using older schema introduced two defects: imposed occupied vacant invariant and incomplete use of PLB
- Implemented fixes for two errors found during September PPMF release

Began executing Sprint VIII

- Science updates implemented during this Sprint will be incorporated in the next PPMF and Detailed Summary Metrics for public release
- Release anticipated in November, to allow time for feedback prior to DSEP meetings scheduled in December

Production Readiness Review (PRR) status

- Delivered solution requirements for Minimum Viable Product (MVP)
- Testing requirements that perform error reporting
- Developing near-term solution for the noisy measurement file format and storage – will complete prior to Operational Readiness Review (ORR)
- Revised 10 of 11 SDLC documents and uploaded to the TI repository
  - DITD will produce the Test Analysis Report (TAR)

Group I Status

Challenges and Issues

Continue to make improvements in the DAS for the production run, such as improvements to system security (e.g., removing floating-point reliance from primitives)

May need to make decision about invariants using development code

Discussion Topic: DAS Schedule Updates: Re-issue of PPMF (vintage 20200917)

Discussion Topic: Recommendations for accuracy/privacy for PL 94-171 from Census Redistricting & Voting Rights Data Office (CRVRDO)

Group I Status

Discussion Topic: DAS Schedule Updates

Multiple activities required to produce re-run of MDFs to re-issue PPMF (vintage 20200917), including improved automated quality control

Planning is in progress to set a target release for the updated PPMF in early November

DSEP meeting to finalize invariants now November 19, 2020

Run experiments for PL94-171/DHC privacy-accuracy trade-offs in early December

Need to shift December DSEP decisions to set PLB forward to January

Group I Status

Discussion Topic: Privacy/Accuracy for PL94-171

Recommendations for accuracy/privacy for PL 94-171 from Census Redistricting & Voting Rights Data Office (CRVRDO)

2020 DAS: Group I Status

30 Day Outlook

DSEP Meeting to set invariants re-scheduled for November 19th

Continue to execute against Sprint VIII executive priorities

Prepare for Operational Readiness Review (ORR)

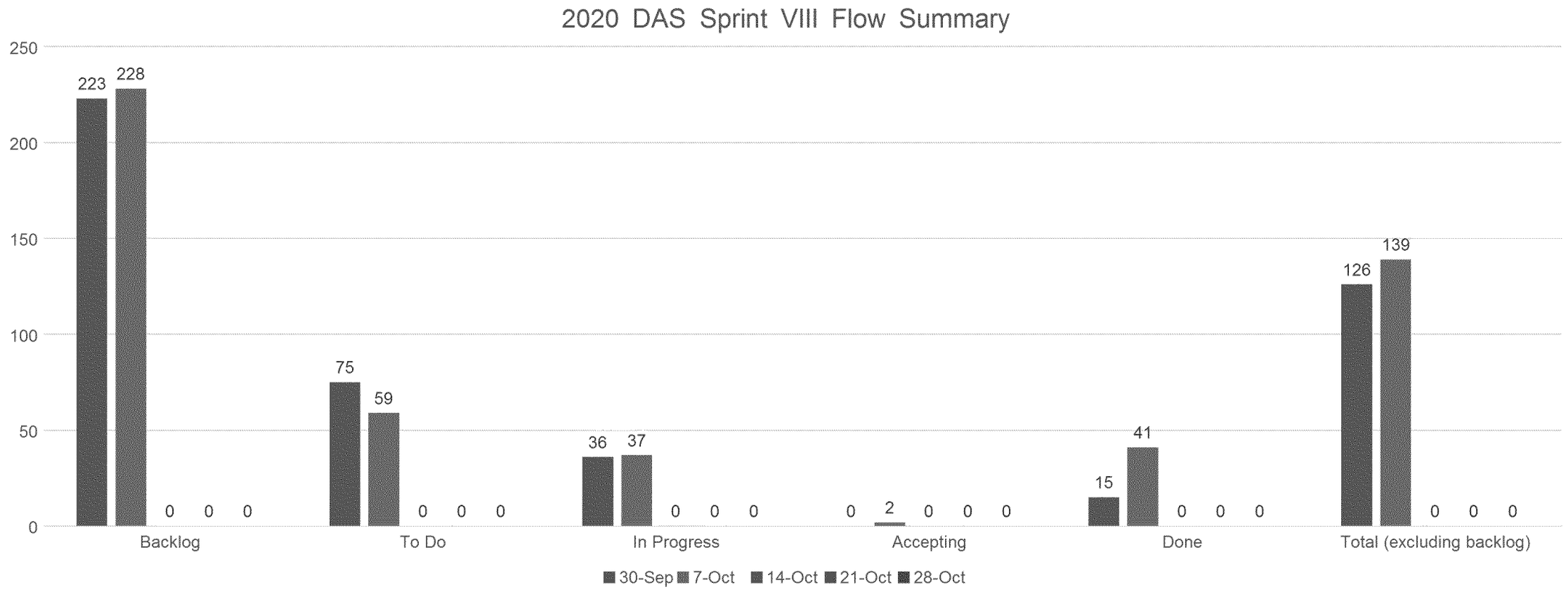
Develop communications plan to explain major functionality contained in each Sprint to external data users

Build on quality assurance/control process for data products produced

Finalize timing of additional Detailed Summary Metrics and PPMF – both for setting invariants and for making PLB decision

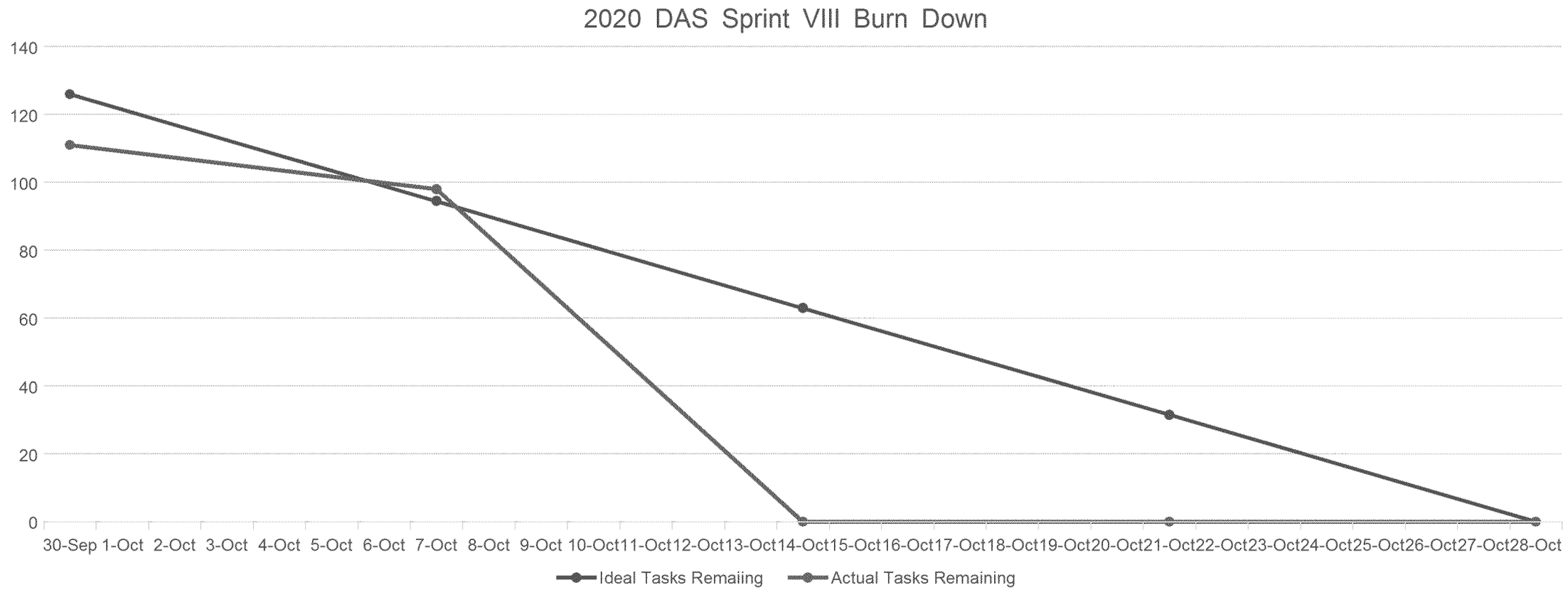
Sprint VIII Flow Summary

Status as of 10/5/2020



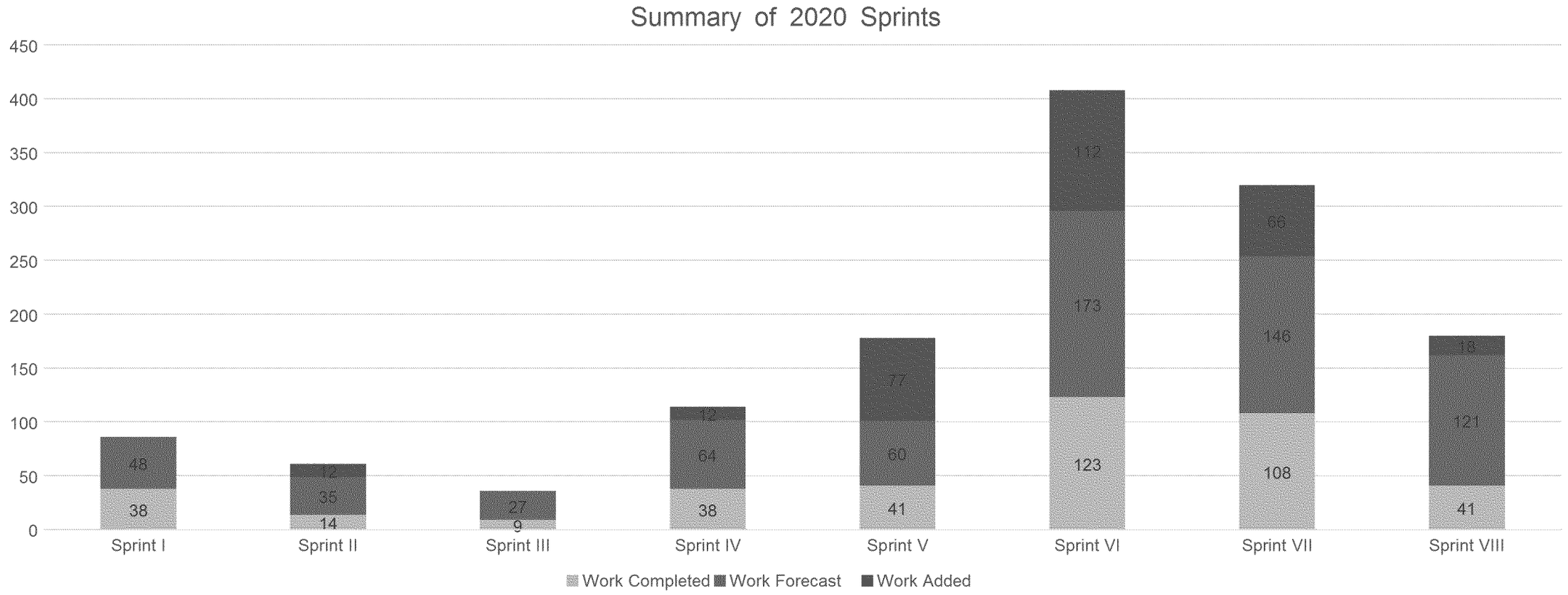
Sprint VIII Burn Down

Status as of 10/5/2020



Summary of 2020 Sprints

Status as of 10/5/2020

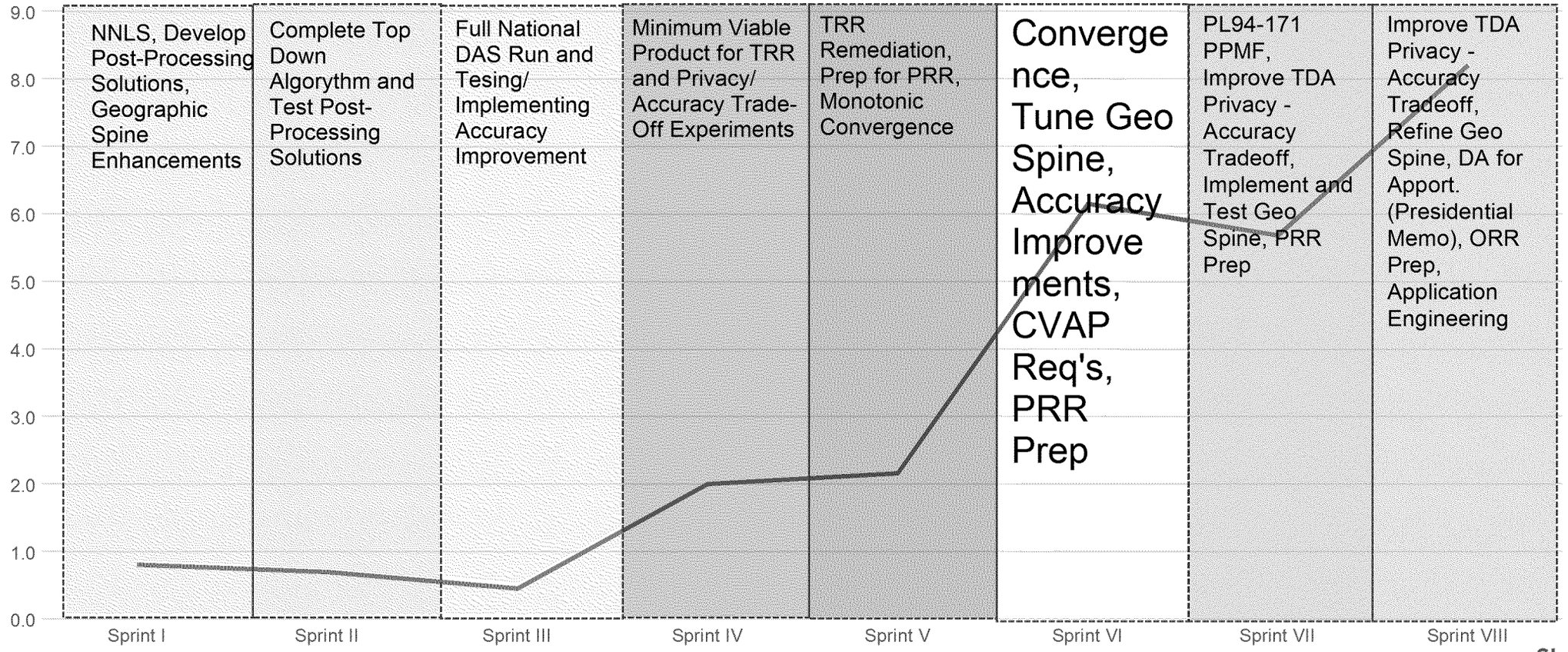


2020 DAS Sprints – Velocity

Status as of 10/5/2020

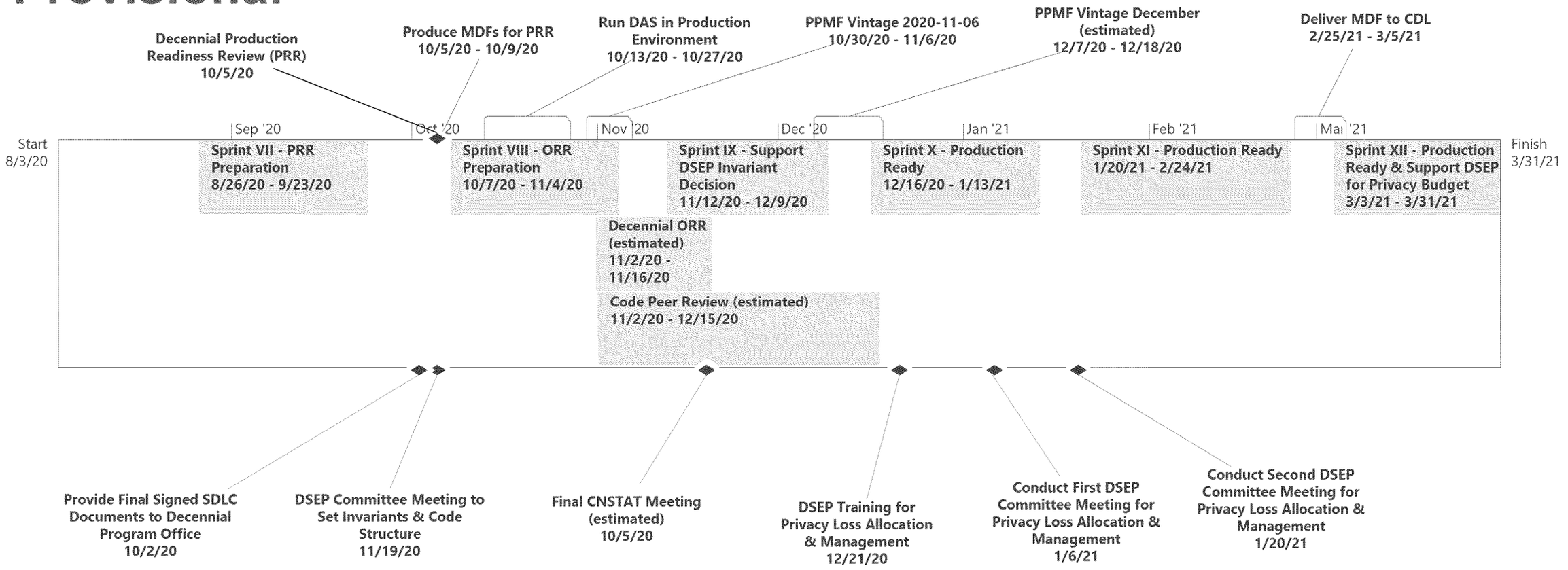
Velocity - Issues Completed per Day

(dotted line moving average)



2020 DAS: Group IA Milestones

Provisional



Milestones continue to be adjusted based on Decennial re-planning effort to meet statutory deadlines

Group II Data Products

2020 DAS

2020 DAS: Group II Status

Accomplishments to Date

Group II Data Products consist of American Indian, Alaska Native (AIAN); Detailed Race and Hispanic Origin; and Person-Household Joins and Averages

Tumult is developing the Group II Products under a contract to MITRE

Outlined initial work plan for year two of the contract period, including major high-level tasks and a tentative schedule

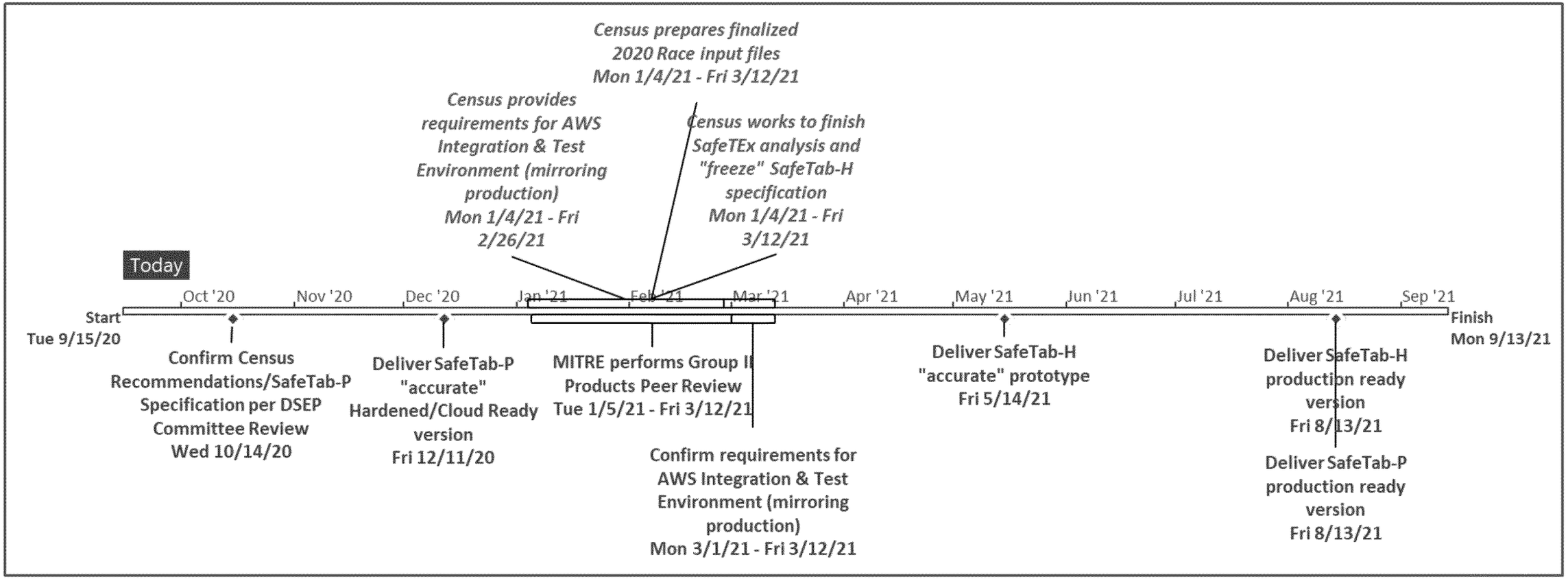
Held working session to discuss the Person-Household Joins

Tumult delivered Simple Prototype SafeTab-H and Cloud Ready Simple SafeTab-P Prototype

Draft Timelines for FY21 Group II Data Products

(as of September 24, 2020)

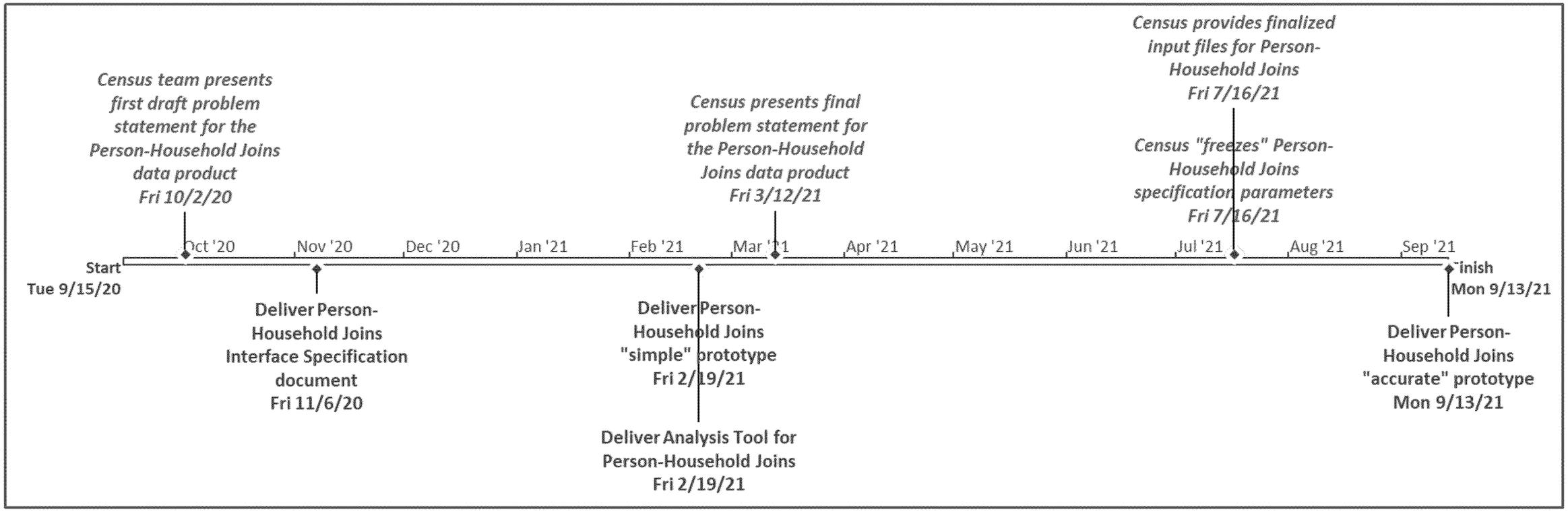
Group II DHC Products (SafeTab-P and SafeTab-H)



- Check mark = completed activity
- All activities performed by a combination of Tumult, MITRE and Census
- Census milestones are in *blue italics*

Draft Timelines for FY21 Group II Data Products
(as of September 24, 2020)

Person-Household Joins



- Check mark = completed activity
- All activities performed by a combination of Tumult, MITRE and Census
- Census milestones are in *blue italics*

2020 DAS: Group II Status

Challenges and 30 Day Outlook

POP delivers research paper, which specifies the capabilities of the SafeTEx product

- October DSEP Meeting to review paper and the epsilon needed for acceptable data quality

Create an operational environment for Tumult SafeTab

- We need to have an operational environment to accept the program and for coordination with other teams (TAB, CEDSCI)
- Implement SafeTab-P and SafeTab-H prototypes in our environment

Execute against timeline for contract year 2

Impact of 2020 operational schedule

Program Communications and Management Activities

2020 DAS

2020 DAS: Communications & Outreach

Online Communication

- **RM Blog** (8/20)
- **DAS Newsletters** (9/18, 9/25)
- **Release of PPMF** (9/17)
- **Release of Detailed Summary Metrics** (9/24)

Expert Groups

- **CNSTAT**
  - *Full expert group (6/1)*
  - *5 working groups (ongoing)*
- **NAC/CSAC**
  - *CSAC presented their first set of recommendations on 9/18*
  - *NAC meeting (9/21)*

Stakeholder Outreach

- **Civil Rights Groups**
  - *Meeting weekly*
- **Census Counts Coalition** (9/11)
- **Washington SDC** (10/6)
- **Oregon SDC** (10/7)
- **CDC DP seminar series** (beginning in October)
- **AIAN Tribal Consultations** (planning for November)

Correspondence/Oversight

- **GAO 104145 (meeting bi-weekly)**
- **FOIA Request Zhou DOC-CEN-2020-001408**
- **CQAS-10523 (Congressional Native American Caucus)**
- **CQAS-10591 (Maine SDC)**
- **CQAS-10611 (30 members of Congress)**

Starting work on Differential Privacy handbook

2020 DAS: Product Release Schedule

	Data Product	2010	2020*
Group IA	PL94-171 (Redistricting Data)	2/3/2011 - 3/24/2011 and 4/14/2011[1]	By March 31, 2021
	CVAP	N/A	By March 31, 2021
Group IB	Demographic Profile	5/5/2011 - 5/26/2011	TBD due to COVID19 delays
	Demographic and Housing Characteristics (DHC) File (formerly Summary File 1)	6/16/2011 - 8/25/2011	TBD due to COVID19 delays
Group II	Summary File 2 Successor (Detailed DHC)[3]	12/15/2011 - 4/26/2012	The information contained in the Summary File 2 successor (Detailed DHC) and the AIAN Summary File and the Person/Household Joins will be delivered as one file (name is TBD). TBD date due to issuance of follow-on contract
	American Indian and Alaska Native (AIAN) Summary File[3]	12/13/2012	
Group III	Public Use Microdata Sample (PUMS)	11/12/2014	On hold pending analysis of Group I and II products.
	Congressional District Demographic and Housing Characteristics File	4/11/2013 and 10/19/2017[2]	Spring 2023
	Census Briefs	3/24/2011 - 9/27/2012	December 2020 – Summer 2023
	Population and Housing Tables	9/27/2011 - 11/26/2013	Fall 2021 – Spring 2023
	Special Reports	9/27/2012 - 12/10/2012	Fall 2022 – Winter 2022

1. The National Summary file released on April 14, 2011.
2. The 113th Congressional District File was released on 4/11/2013 and the 115th Congressional District File was released on October 19, 2017. There was no file released for the 114th Congressional District as the states reported there were no changes for their boundaries.
3. Person/household joins and averages are part of this product.

***Planning for the production and release of the remaining 2020 Census data products will restart immediately following completion of the apportionment and redistricting data planning activities.**

2020 DAS: Budget FY21 Annual Plan

Division	Project Code	Annual Plan	Plan to Date	Expended to Date	Variance to Date
90 – ADRM	6650F08	$260,752	$0.00	$0.00	$0.00
90 – ADRM	6750F95	$306,942	$0.00	$0.00	$0.00
92 – CED	6650F08	$6,243,417	$0.00	$0.00	$0.00
92 – CED	6750F95	$36,534	$0.00	$0.00	$0.00
Totals		$6,847,645	$0.00	$0.00	$0.00

2020 DAS: Program/Project Risks

Program Risks	Risk Status
DPD004 – 2020 DAS is Employing a Modern Privacy Protection	• Continue to execute project to demonstrate DAS capability to protect confidential data while providing data that is fit for use.
DPD014 – Key Personnel Are Not Replaceable	• Management continues to work with DAS team to identify required resources and coordinate with Census organizations to acquire
Project Risks	**Risk Status**
DAS015 – Parameters to Support Differential Privacy (in Addition to Epsilon)	• Conduct training for DSEP in order to support their required input on invariants and establishment of the privacy loss budget (i.e., epsilon).
DAS024 – Test 2020 CEF	• A final version of the format of the 2020 CEF file will be created. However, it is imperative that this file format and test file based on this format be provided in time to allow for the DAS to test its capability to use. • Working with DEMO to determine when the 2020 CEF file can be provided and currently using 2010 CEF file for testing.
DAS025 - Algorithm Effectiveness for 2020 DAS Group II products	• During testing of the Group II data products an additional assessment will be done as part of the data quality check and will determine if PII can be re-identified
DAS026 - Expectation for SafeTab-P and SafeTab-H prototypes and final products to meet delivery schedule	• Continue to work with contractor team and Demo to define and refine requirements so that an early warning system exists for potential shifts in executive priorities, including from the Department and to minimize any misunderstanding of the requirements

Appendix

2020 DAS Scope, Schedule Planning

2020 DAS: Scope – Data Products

Group IA & IB	Group II	Group III
• PL94-171 • Demographic Profiles • Demographic and Housing Characteristics-Persons (DHC-P) • DHC-Households (DHC-H) ❖ Citizen Voting Age Population (CVAP)	• American Indian, Alaska Native (AIAN) • Detailed Race and Hispanic Origin • Person-Household Joins and Averages	• Public Use Microdata Sample (PUMS) • Other Special Tabulations • Noisy Measurements File Format

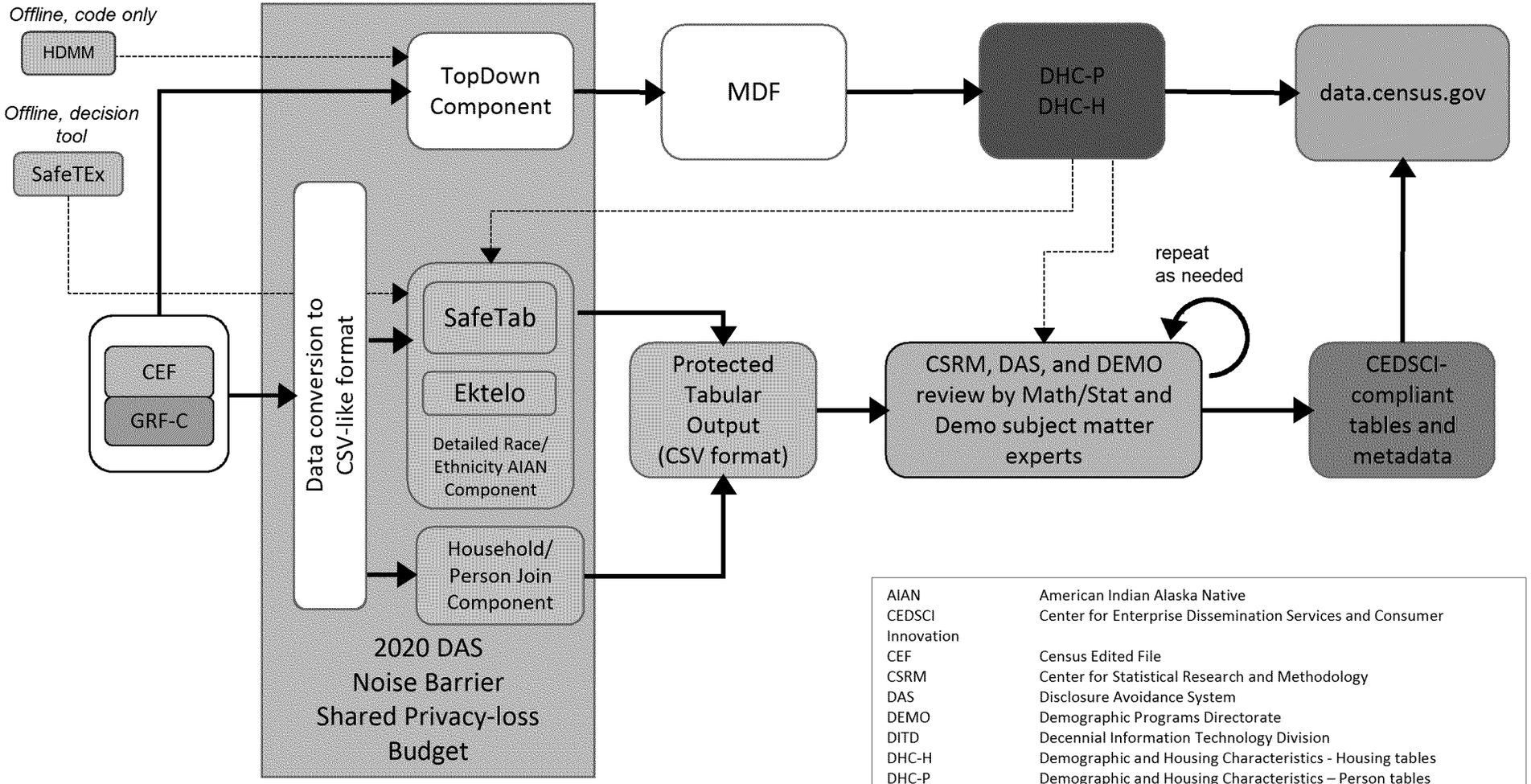
❖ The CVAP is a special tabulation:

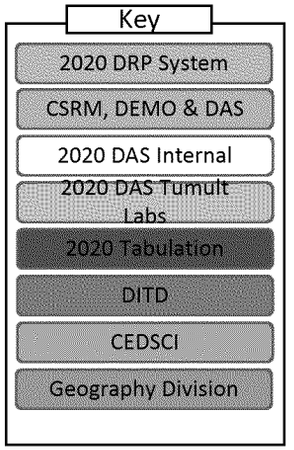
- ➢ Will not be available for disclosure avoidance processing until after the Microdata Detail File (MDF) has been finalized, on *TBD*.
- ➢ The protected CVAP data must be produced and released to tabulation by *TBD*.
- ➢ Consistency with PL94-171 tabulations is required and will include race/ethnicity categories to be constructed by collapsing cells in the P4 table of PL94-171.

2020 DAS: Scope

Component Chart for 2020 DAS

Version 8.6
2020-04-21





AIAN	American Indian Alaska Native
CEDSCI Innovation	Center for Enterprise Dissemination Services and Consumer
CEF	Census Edited File
CSRM	Center for Statistical Research and Methodology
DAS	Disclosure Avoidance System
DEMO	Demographic Programs Directorate
DITD	Decennial Information Technology Division
DHC-H	Demographic and Housing Characteristics - Housing tables
DHC-P	Demographic and Housing Characteristics – Person tables
DRP	Decennial Response Processing
GRF-C	Geographic Reference File-Codes
MDF	Micro-data Detail File

Sprints Planned for 2020 DAS Production

Decennial Re-plan Schedule

Sprint #	Sprint Planning	Sprint Execution	Inter-sprint	High-level Priorities
VII	August 24	August 26 – September 23	September 23 – 29	PRR Prep, Support DSEP invariants decision
Production Readiness Review (PRR): October 5, 2020				
VIII	September 28	September 30 – October 28	October 28 – November 3	Bug Fixes. Tuning via the config file. ORR prep. PPMF production.
Operational Readiness Review (ORR): TBD November 2020				
IX	November 2	November 4 – December 2	December 2 – December 8	Support for DSEP decision for PLB. CVAP development
Presidential Memo Special Tabulation: December 31, 2020				
X	December 7	December 9 – Jan 6 (use/lose)	January 6 – January 12	Production Ready for PL (Group IA). Runtime environment. No TDA changes
XI	January 11	January 13 – February 10	February 10 – 16	2020 DAS Production Run prep
Produce Privacy-Protected 2020 Census Microdata Detail File (MDF): TBD				
XII	February 15	February 17 – March 17	March 17 – 23	2020 DAS Production Run Group IB* for DHCs
PL94-171 Redistricting File & CVAP Special Tabulation: originally March 31, 2021				
XIV	March 22	March 24 – April 21	April 21 – 28	

*not exactly as currently designed

Application Engineering & DevOps

Goals for 2020 DAS

- **Goal 1:** Assure the reliable operation of the 2020 DAS for the PL94-171 production runs.
- **Goal 2:** Create a system that allows the science team to productively complete its work.
- **Goal 3:** Control costs by using Amazon Web Services as efficiently as possible.
- **Goal 4:** Support good systems development practices by continuously reviewing and updating the DAS SDLC documentation.

EXHIBIT 19

The U.S. Census Bureau Tries to Be a Good Data Steward in the 21st Century

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
Privitar In:Confidence USA
Wednesday, November 14, 2019 11:35-12:05pm

United States Census Bureau

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

The views expressed in this talk are my own and not those of the U.S. Census Bureau. Statistics from the 2010 Census for Rhode Island authorized under DRB release CBDRB-FY19-054.

Acknowledgments

The Census Bureau's 2020 Disclosure Avoidance System incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Computer Scientist), Rob Sienkiewicz (Chief, Center for Enterprise Dissemination), Tamara Adams, Robert Ashmead, Craig Corl, Jason Devine, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Edward Porter, Sarah Powazek, Anne Ross, Ian Schmutte, William Sexton, Victoria Velkoff, Lars Vilhuber, and Pavel Zhuralev

The challenges of a census:

1. collect all of the data necessary to underpin our democracy
2. protect the privacy of individual data to ensure trust and prevent abuse

Major data products:

- Apportion the House of Representatives
  (due December 31, 2020)
- Supply data to all state redistricting offices
  (due April 1, 2021)
- Demographic and housing characteristics
  (no statutory deadline, target summer 2021)
- Detailed race and ethnicity data
  (no statutory deadline)
- American Indian, Alaska Native, Native Hawaiian data
  (no statutory deadline)

For the 2010 Census, this was *more than 150 billion* statistics from 15GB total data.

Estimate: 15GB of final data from 2020 Census ($1/byte!)

Less than 1% of worldwide mobile data use/second

(Source: Cisco VNI Mobile, February 2019 estimate: 11.8TB/second, 29EB/month, mobile data traffic worldwide https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html#_Toc953327.)

The Census Bureau's data stewardship problem looks very different from the one at Amazon, Apple, Facebook, Google, Microsoft, Netflix, Uber, …

… but appearances are deceiving.

The Database Reconstruction Vulnerability

What we did

- Database reconstruction for all 308,745,538 people in 2010 Census
- Link reconstructed records to commercial databases: acquire PII
- Successful linkage to commercial data: putative re-identification
- Compare putative re-identifications to confidential data
- Successful linkage to confidential data: confirmed re-identification
- Harm: attacker can learn self-response race and ethnicity

What we found

- Census block and voting age (18+) correctly reconstructed in all 6,207,027 inhabited blocks
- Block, sex, age (in years), race (OMB 63 categories), ethnicity reconstructed
  - Exactly: 46% of population (142 million of 308,745,538)
  - Allowing age +/- one year: 71% of population (219 million of 308,745,538)
- Block, sex, age linked to commercial data to acquire PII
  - Putative re-identifications: 45% of population (138 million of 308,745,538)
- Name, block, sex, age, race, ethnicity compared to confidential data
  - Confirmed re-identifications: 38% of putative (52 million; 17% of population)
- For the confirmed re-identifications, race and ethnicity are learned correctly, although the attacker may still have uncertainty

Almost everyone in this room knows that:

Comparing common features allows highly reliable entity resolution (“these features belong to the same entity”)

Machine learning systems build classifiers, recommenders, and demand management systems that use these amplified entity records (“features predict outcomes”)

All of this is much harder with provable privacy guarantees for the entities!

Privacy protection is an economic problem,

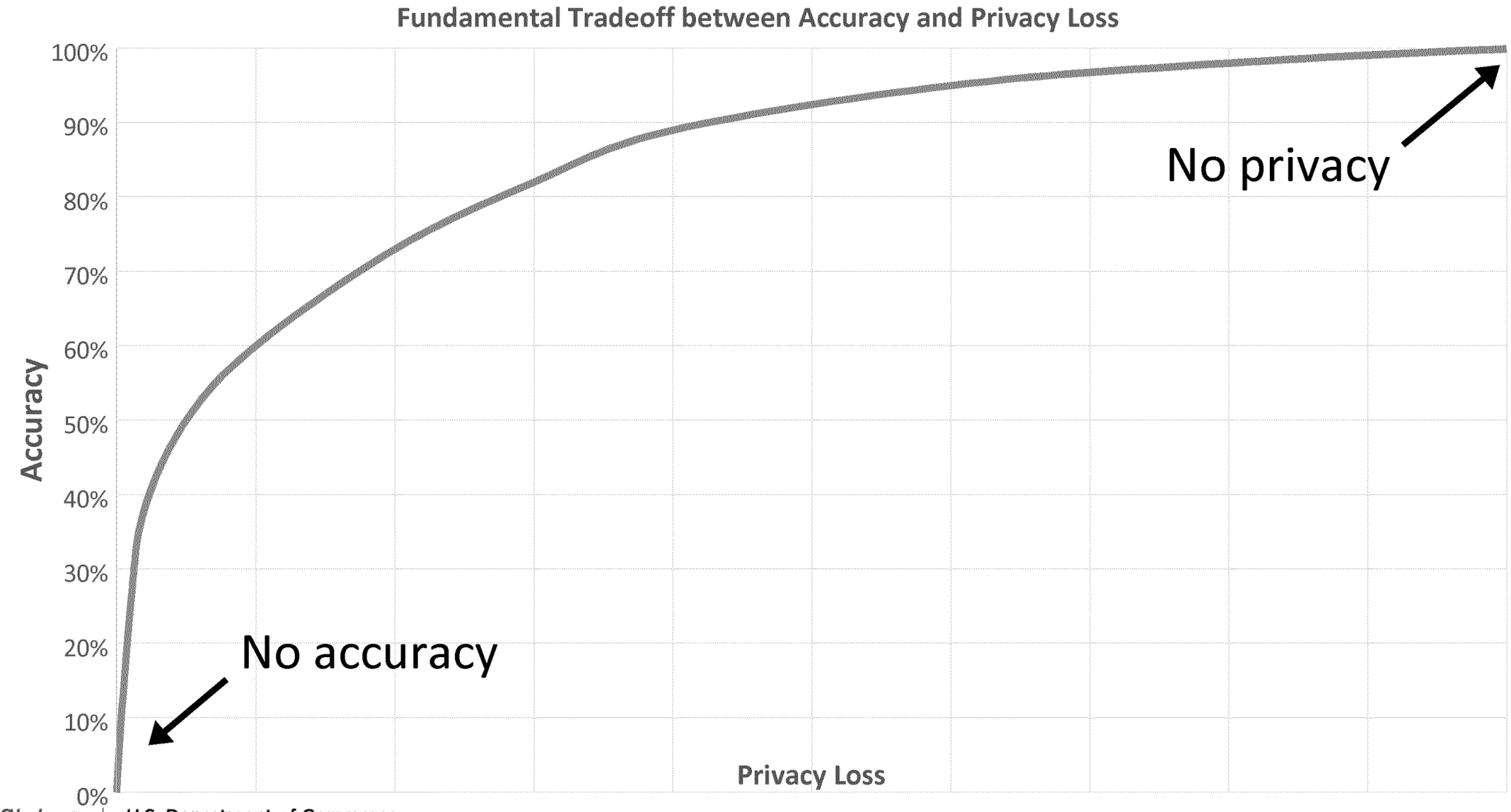
Not a technical problem in computer science or statistics.

Allocation of a scarce resource (data in the confidential database) between competing uses:

information products

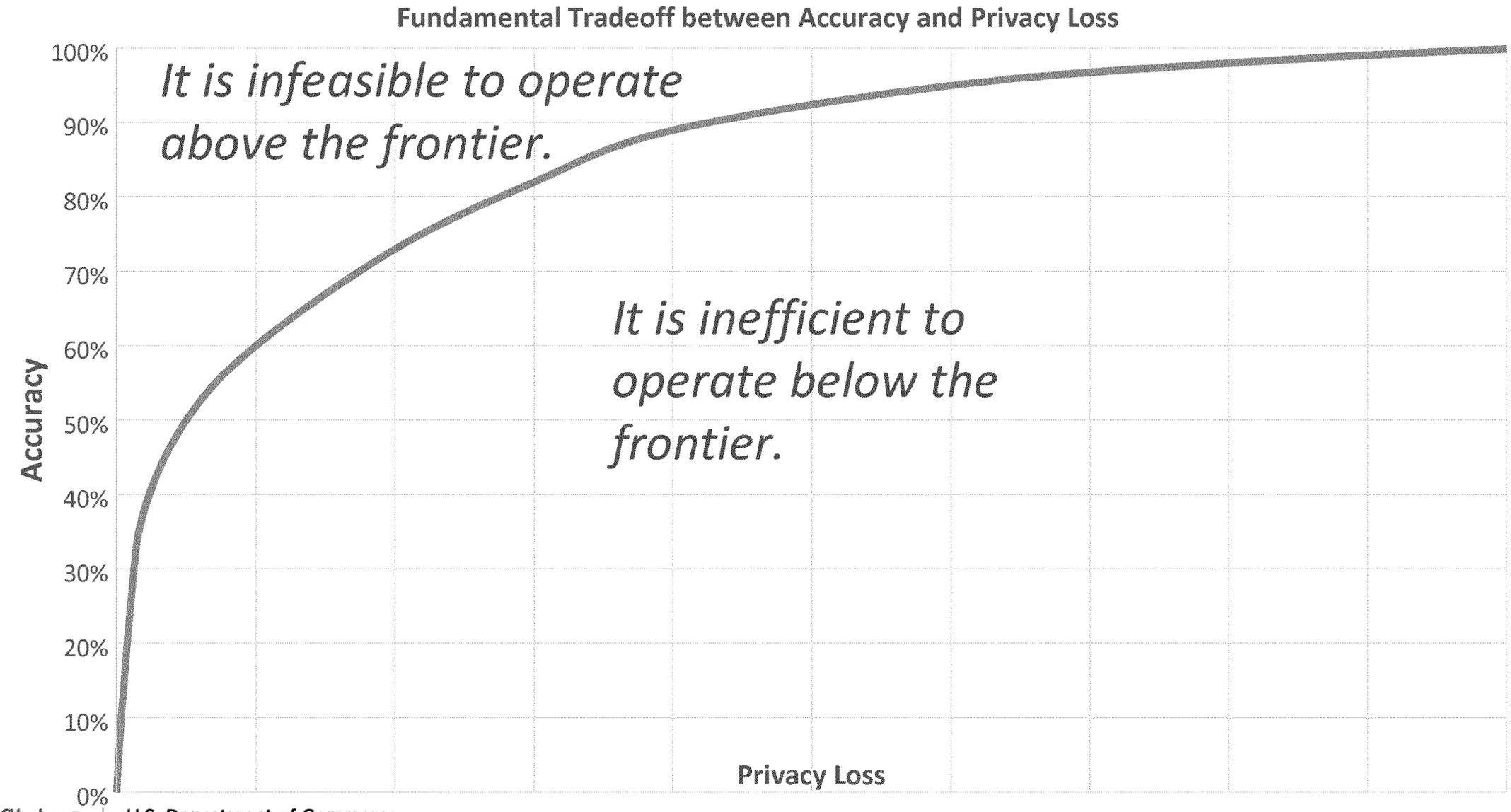
and

privacy protection.



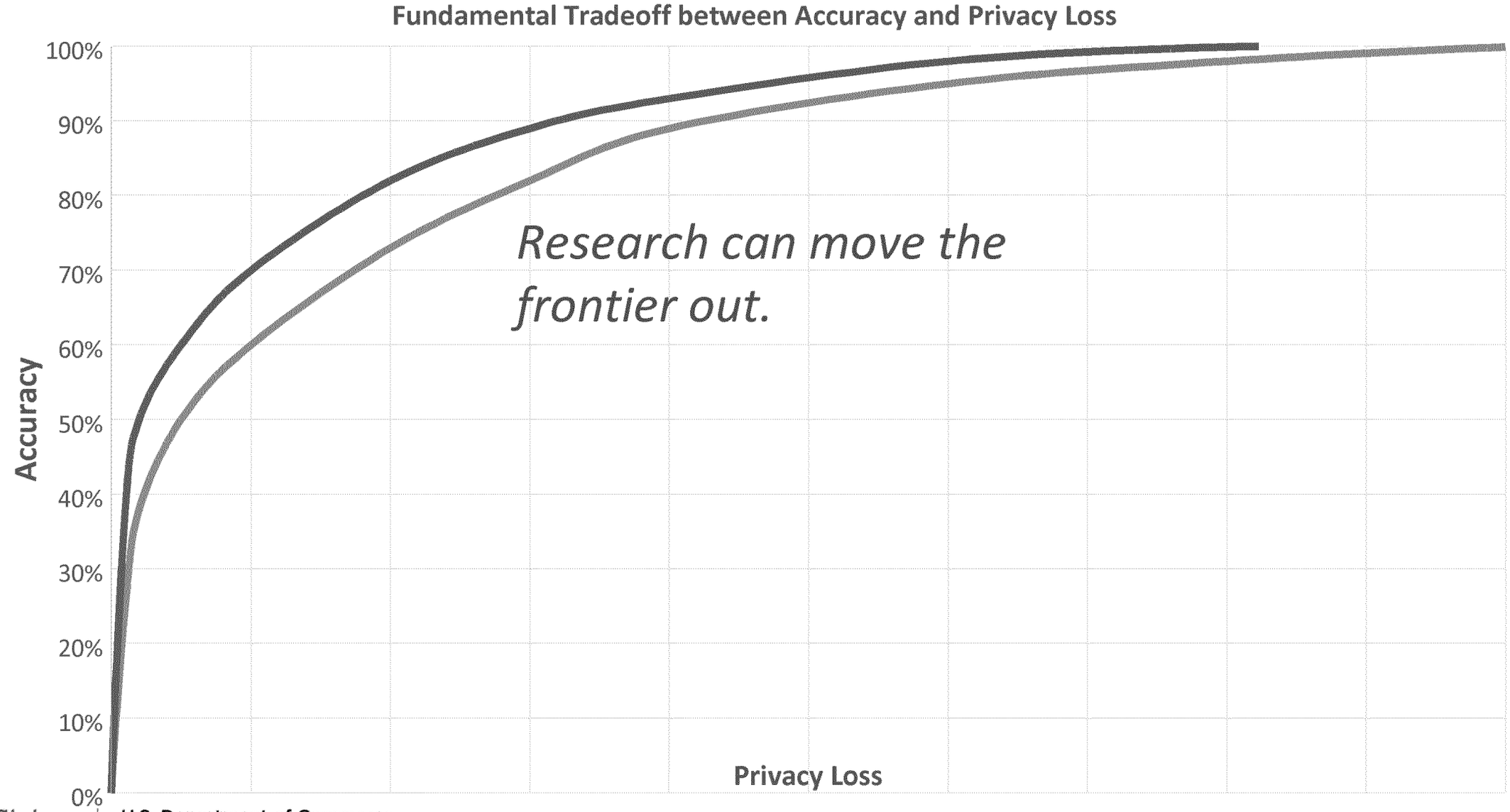
United States Census Bureau

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



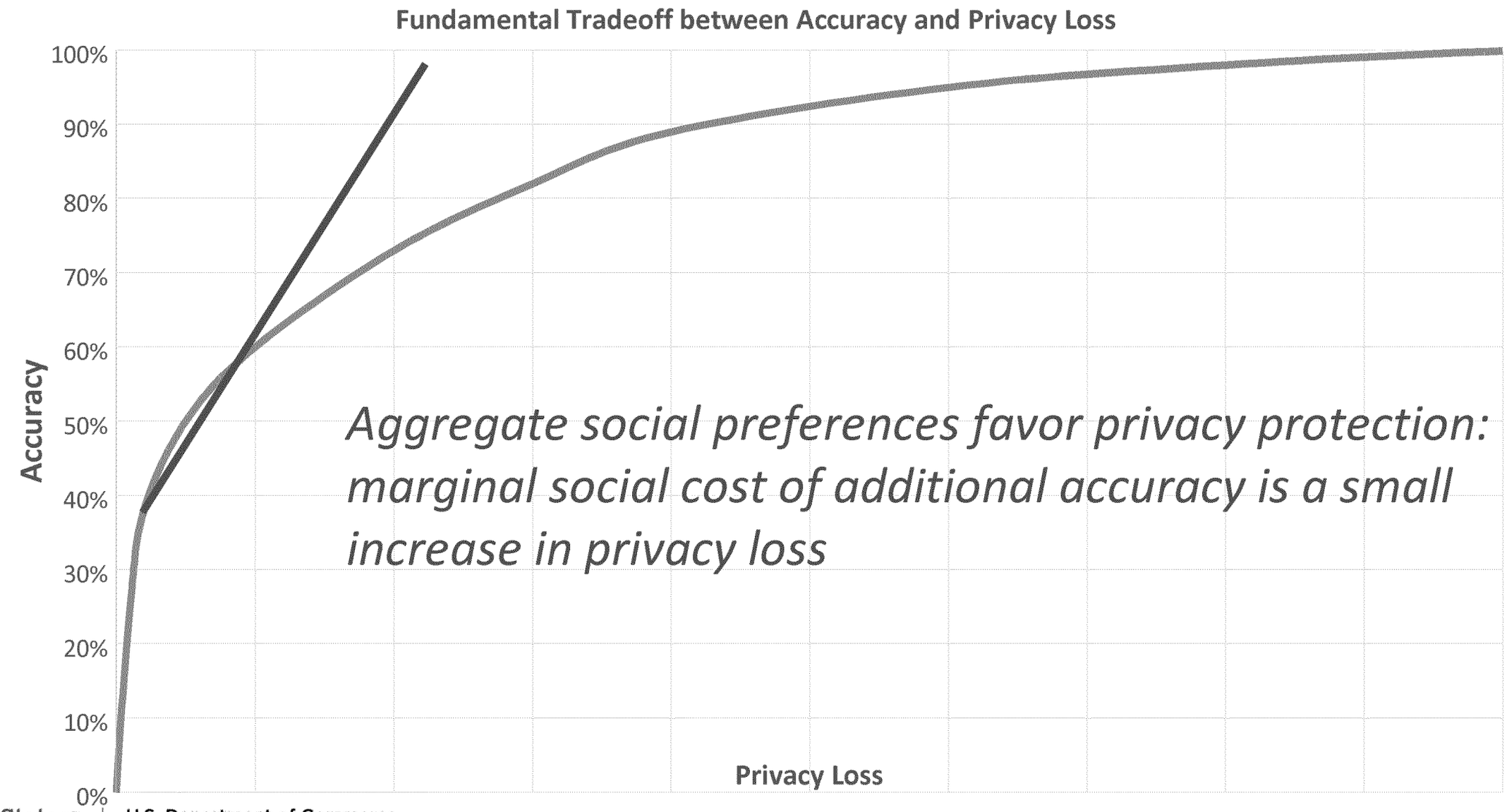
United States Census Bureau

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



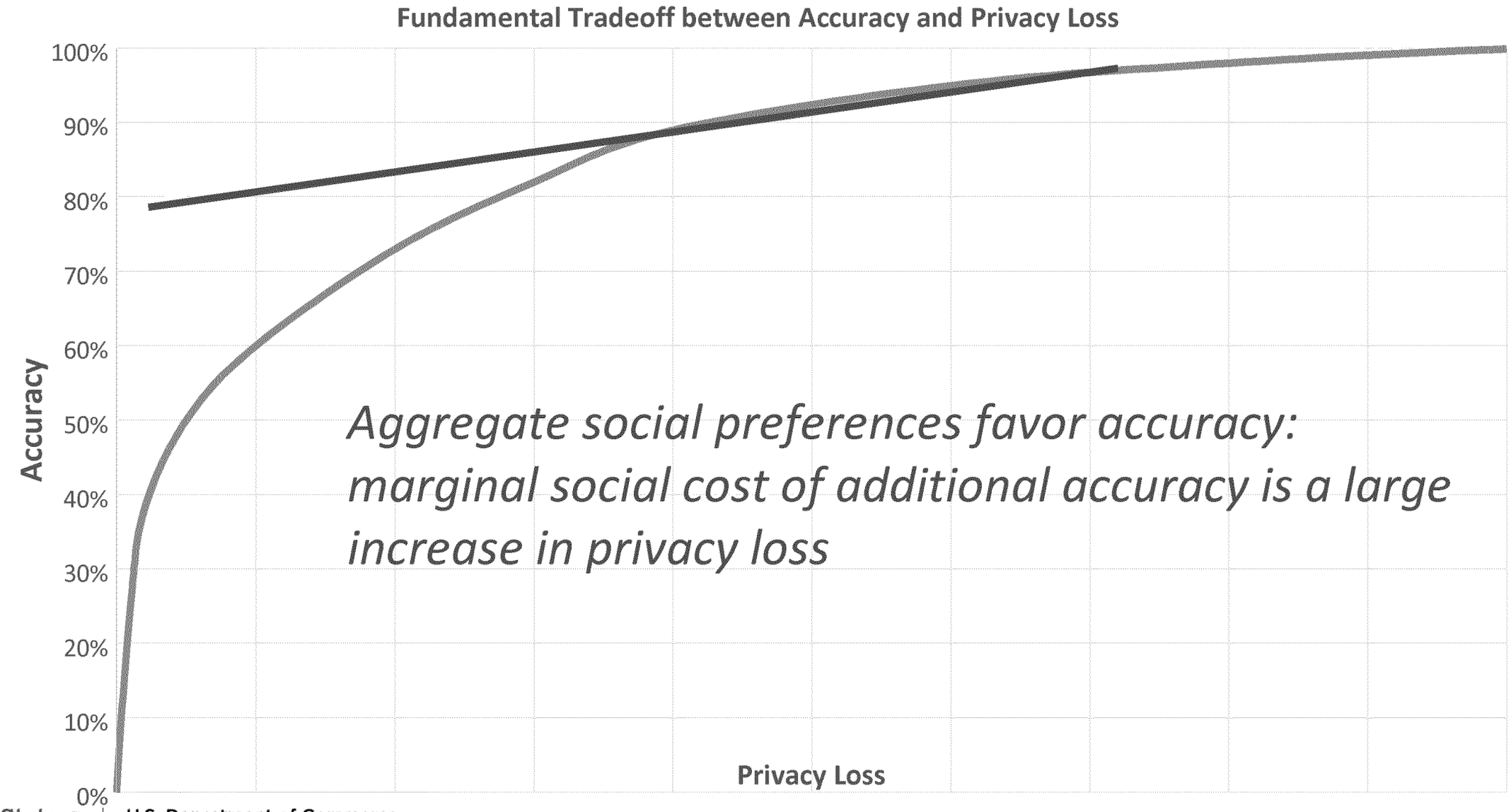
United States™ Census Bureau

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



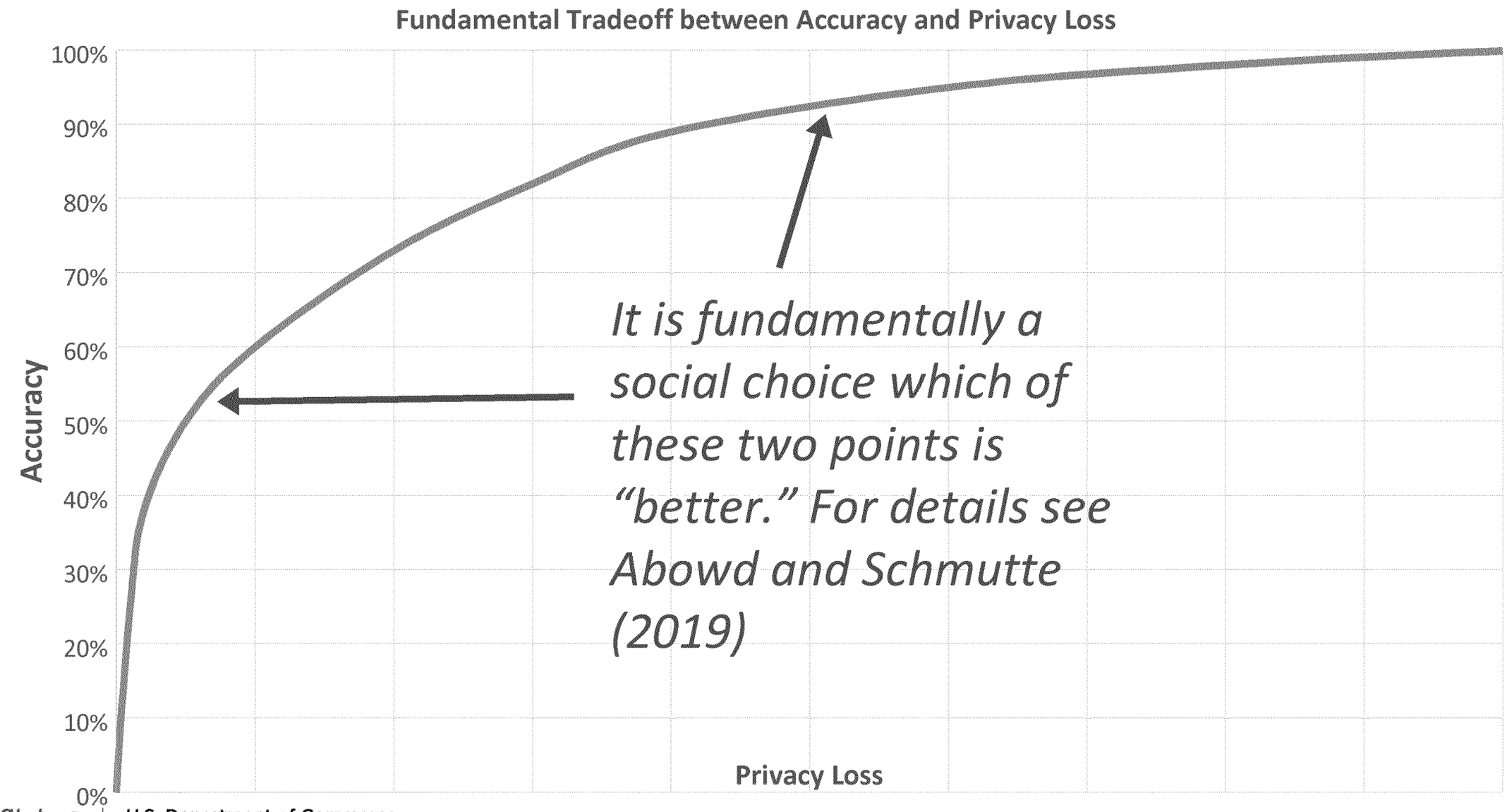
United States™ Census Bureau

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



United States™ Census Bureau

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



United States™ Census Bureau

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

All 2020 Census Publications

- Will all be processed by a collection of differentially private algorithms
- Using a total privacy-loss budget set as policy, not hard-wired, determined by the Data Stewardship Executive Policy Committee
- Code base, technical documents, and extensive demonstration products based on the 2010 Census confidential data have all been released to the public
- More information: https://www.census.gov/newsroom/blogs/research-matters/2019/10/balancing_privacyan.html

Statistical data, fit for their intended uses, can be produced when the entire publication system is subject to a formal privacy-loss budget.

To date, the team developing these systems use ε-differential privacy for the data publications from the 2020 Census used:

- to re-draw every legislative district in the nation (P.L. 94-171 tables)
- to support the bulk of the demographic tables from the former Summary File 1 (now called Demographic and Housing Characteristics).

But there were more than 100 billion other queries published from the 2010 Census that are not easy to make consistent with a finite privacy-loss budget.

The 2020 Disclosure Avoidance team has also developed methods for quantifying and displaying the system-wide trade-offs between the accuracy of the decennial census data products and the privacy-loss budget assigned to sets of tabulations.

Considering that work began in mid-2016 and that no organization anywhere in the world has yet deployed a full, central differential privacy system, this is already a monumental achievement.

Algorithms Matter

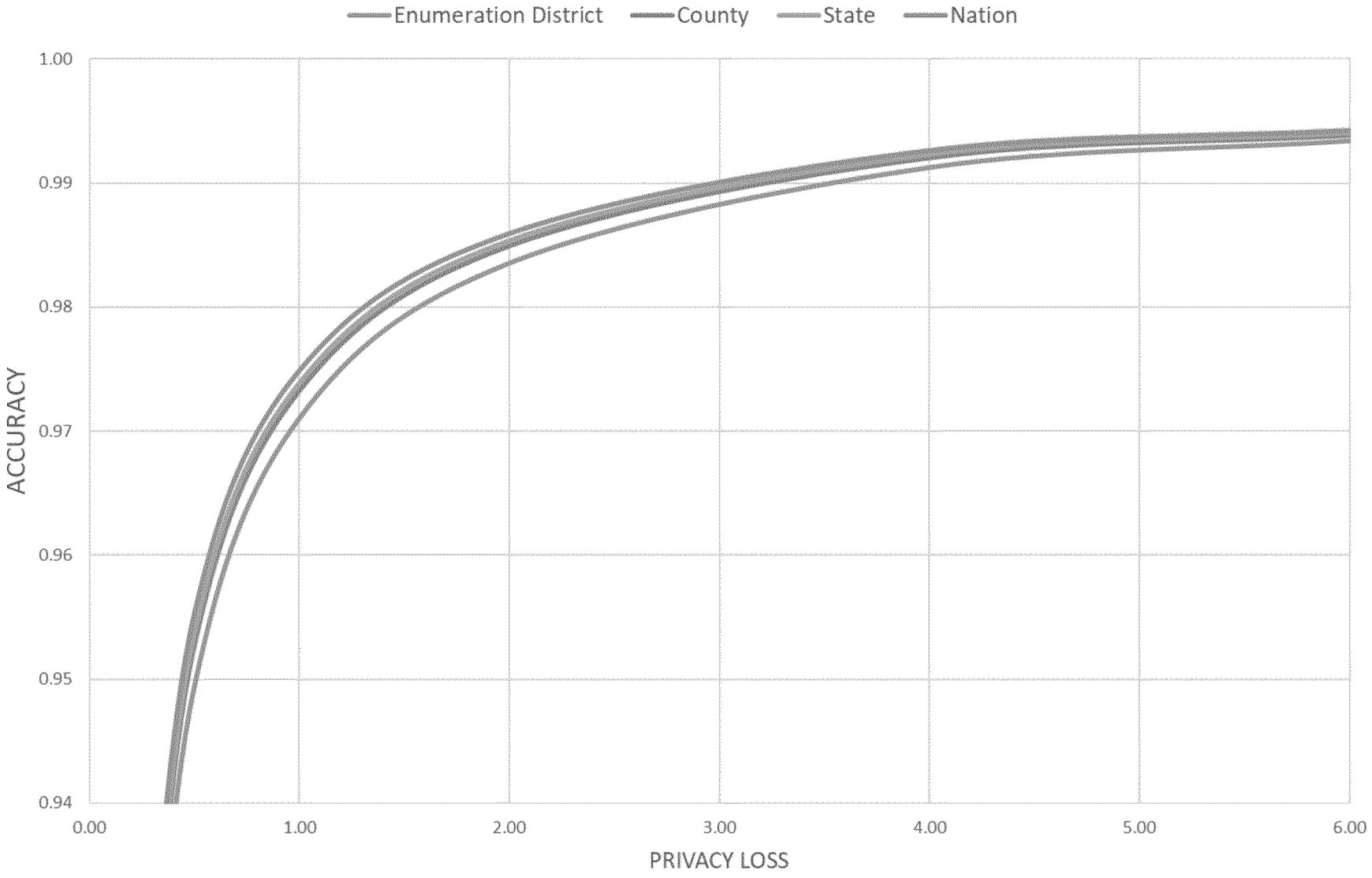
Naïve Method: BottomUp or Block-by-Block

- Apply differential privacy algorithms to the most detailed level of geography
- Build all geographic aggregates from those components as a post-processing
- This is similar to the local differential privacy implementations in the Chrome browser, iOS, and Windows 10.

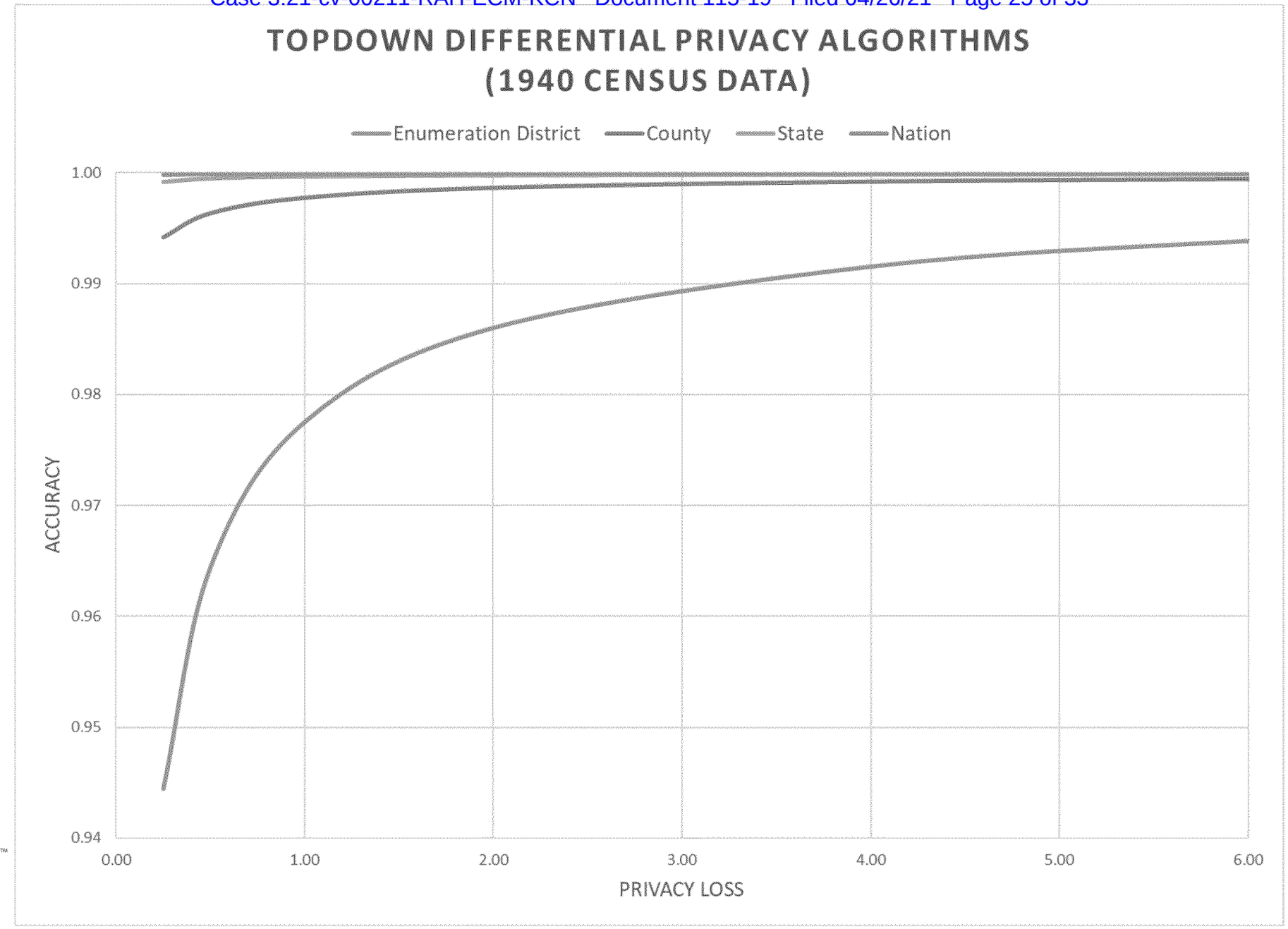
The Census TopDown Algorithm (TDA)

- Take differentially private measurements at every level of the Census geographic hierarchy
- At each level of TDA post-process:
  - Solve an L2 optimization to get non-negative tables
  - Solve an L1 optimization to get non-negative, integer tables
  - Generate micro-data from the post-processed tables

DISTRICT-BY-DISTRICT DIFFERENTIAL PRIVACY ALGORITHMS (1940 CENSUS DATA)

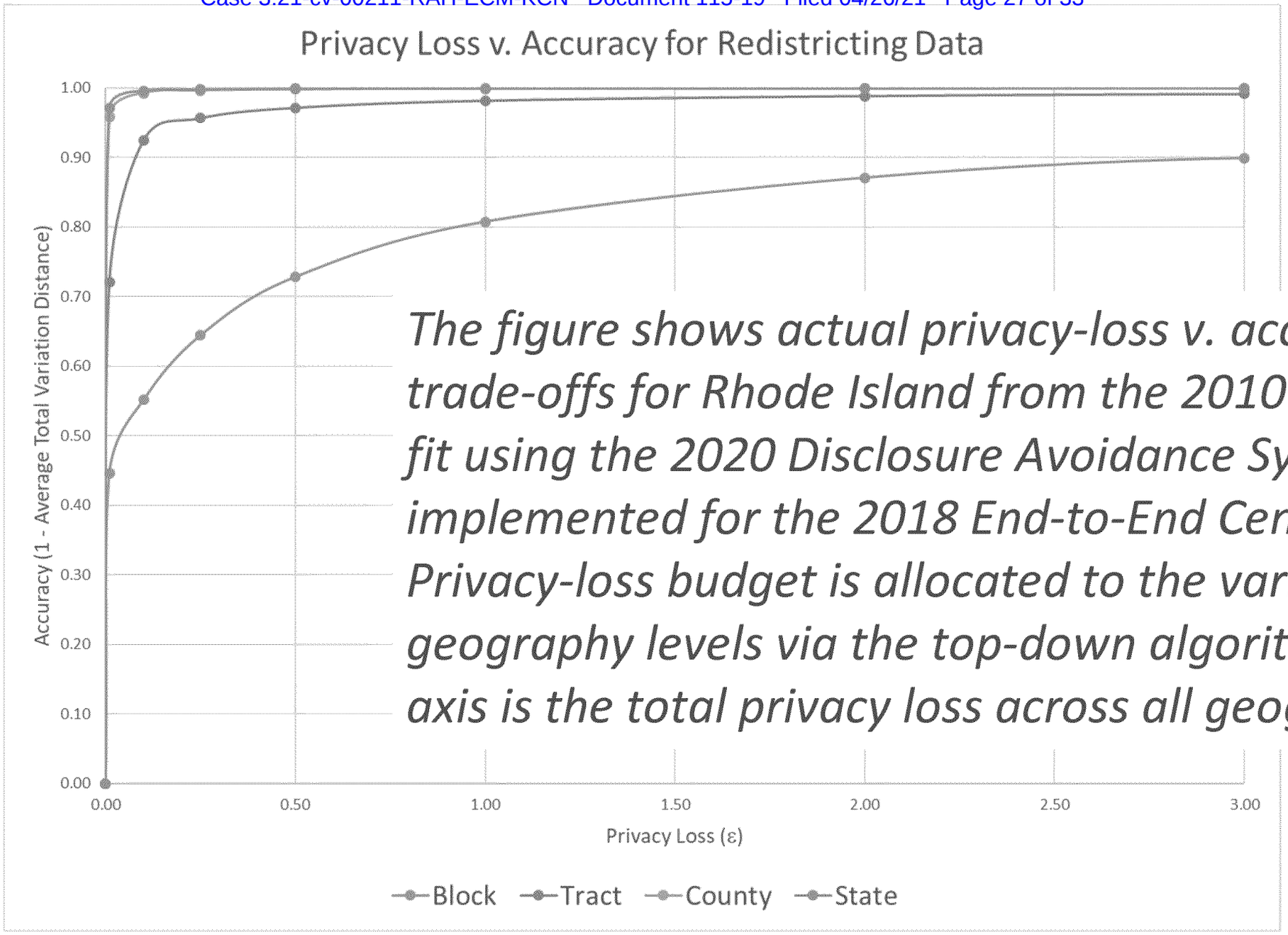


United States™ Census Bureau



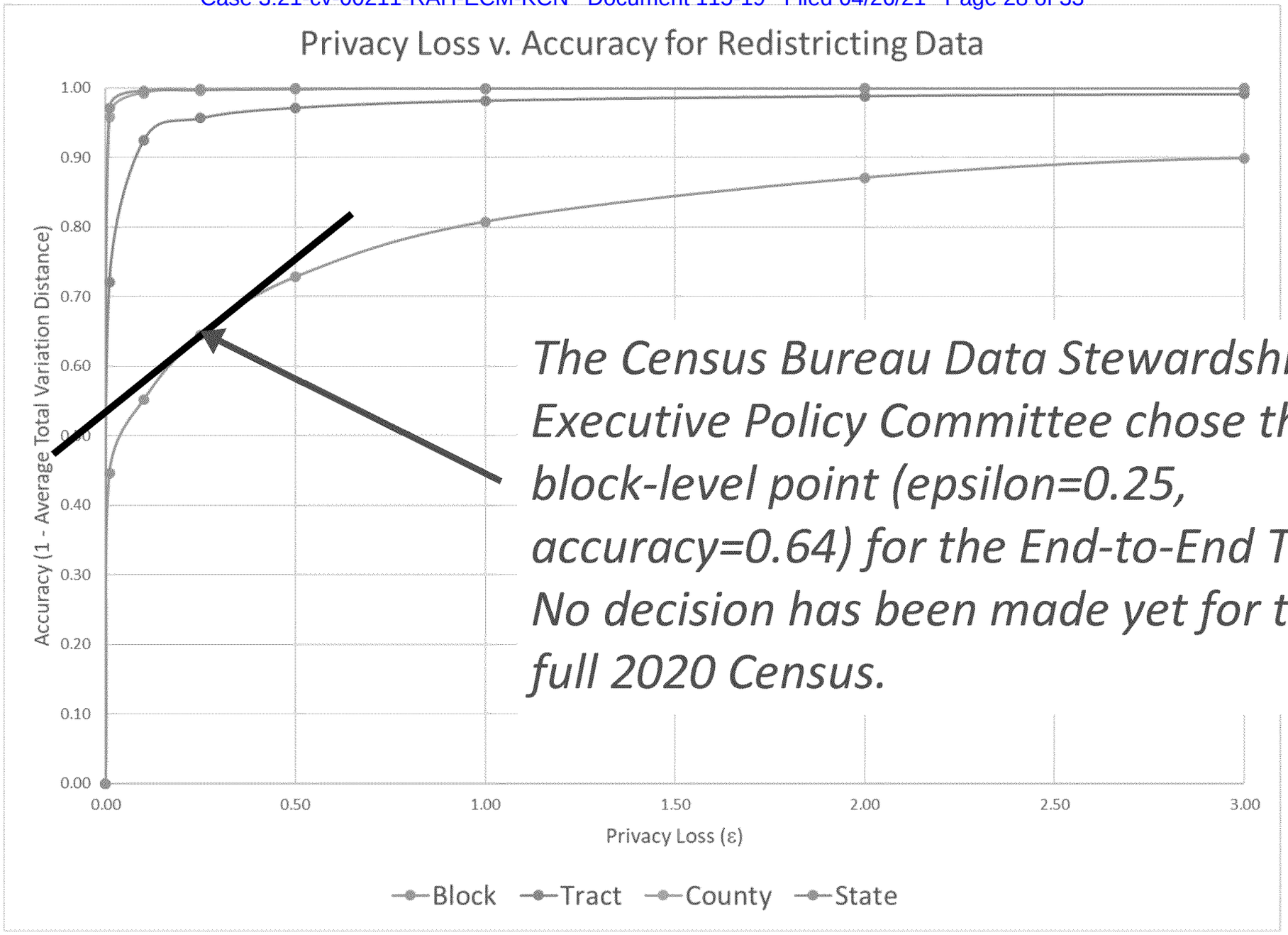
United States™ Census Bureau

Managing the Privacy-Loss Budget



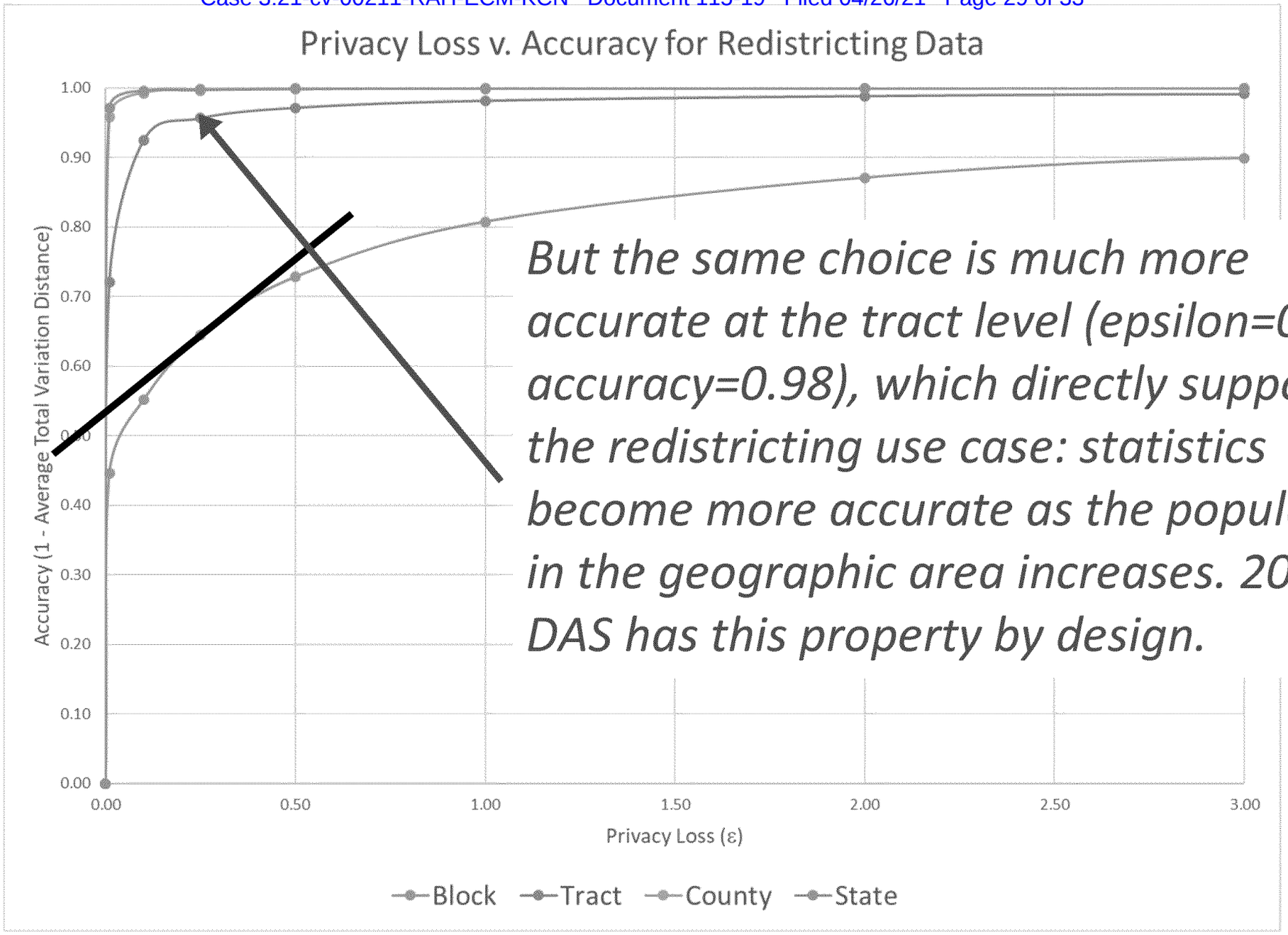
The figure shows actual privacy-loss v. accuracy trade-offs for Rhode Island from the 2010 Census, fit using the 2020 Disclosure Avoidance System as implemented for the 2018 End-to-End Census Test. Privacy-loss budget is allocated to the various geography levels via the top-down algorithm. X-axis is the total privacy loss across all geographies.

United States™ Census Bureau



The Census Bureau Data Stewardship Executive Policy Committee chose this block-level point (epsilon=0.25, accuracy=0.64) for the End-to-End Test. No decision has been made yet for the full 2020 Census.

United States™ Census Bureau



But the same choice is much more accurate at the tract level (epsilon=0.25, accuracy=0.98), which directly supports the redistricting use case: statistics become more accurate as the population in the geographic area increases. 2020 DAS has this property by design.

United States™ Census Bureau

Only the tip of the iceberg

Demographic profiles, based on the detailed tables traditionally published in summary files following the publication of redistricting data, have far more diverse uses than the redistricting data.

Summarizing those use cases in a set of queries that can be answered with a reasonable privacy-loss budget is the next challenge.

Internet giants, businesses and statistical agencies around the world should also step-up to these challenges. We can learn from, and help, each other enormously.

More Background on the 2020 Census Disclosure Avoidance System

- September 14, 2017 CSAC (overall design)
  https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf?#
- August, 2018 KDD'18 (top-down v. block-by-block)
  https://digitalcommons.ilr.cornell.edu/ldi/49/
- October, 2018 WPES (implementation issues)
  https://arxiv.org/abs/1809.02201
- October, 2018 *ACMQueue* (understanding database reconstruction)
  https://digitalcommons.ilr.cornell.edu/ldi/50/ or
  https://queue.acm.org/detail.cfm?id=3295691
- December 6, 2018 CSAC (detailed discussion of algorithms and choices)
  https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#
- June 6, 2019 Blog explaining how to use the 2018 End-to-End Census Test version of the 2020 Disclosure Avoidance System with the 1940 Census public data from IPUMS
  https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html
- October 29, 2019 2010 Demonstration Data Products (blog and links)
  https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html

Selected References

- Dinur, Irit and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*(PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. in Halevi, S. & Rabin, T. *(Eds.)* Calibrating Noise to Sensitivity in Private Data Analysis *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg*, 265-284, DOI: 10.1007/11681878_14.
- Dwork, Cynthia. 2006. Differential Privacy, *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006), Springer Verlag, 4052*, 1-12, ISBN: 3-540-35907-9.
- Dwork, Cynthia and Aaron Roth. 2014. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science. Vol. 9, Nos. 3–4. 211–407, DOI: 10.1561/0400000042.
- Dwork, Cynthia, Frank McSherry and Kunal Talwar. 2007. The price of privacy and the limits of LP decoding. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*(STOC '07). ACM, New York, NY, USA, 85-94. DOI:10.1145/1250790.1250804.
- Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd , Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.
- Dwork, Cynthia and Moni Naor. 2010. On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy, *Journal of Privacy and Confidentiality*: Vol. 2: Iss. 1, Article 8. Available at: http://repository.cmu.edu/jpc/vol2/iss1/8.
- Kifer, Daniel and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (SIGMOD '11). ACM, New York, NY, USA, 193-204. DOI:10.1145/1989323.1989345.
- Abowd, John M. and Ian M. Schmutte. 2019. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*, Vol. 109, No. 1 (January 2019):171-202, DOI:10.1257/aer.20170627.
- Erlingsson, Úlfar, Vasyl Pihur and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (CCS '14). ACM, New York, NY, USA, 1054-1067. DOI:10.1145/2660267.2660348.
- Apple, Inc. 2016. Apple previews iOS 10, the biggest iOS release ever. Press Release (June 13). URL=http://www.apple.com/newsroom/2016/06/apple-previews-ios-10-biggest-ios-release-ever.html.
- Ding, Bolin, Janardhan Kulkarni, and Sergey Yekhanin 2017. Collecting Telemetry Data Privately, NIPS 2017.
- Bittau , Andrea, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Usharsee Kode, Julien Tinnes, and Bernhard Seefeld 2017. Prochlo: Strong Privacy for Analytics in the Crowd, https://arxiv.org/abs/1710.00901.

Thank you.

John.Maron.Abowd@census.gov

EXHIBIT 20

UNITED STATES DEPARTMENT OF COMMERCE
U.S. Census Bureau
Washington, DC 20233-0001

February 26, 2020

Mr. Jeff Hardcastle
State Demographer
4600 Kietzke Lane
Building L, Suite 235
Reno, NV 89502

Dear Steering Committee Members:

Thank you for your letter on the U.S. Census Bureau's adoption of differential privacy to protect the confidentiality of respondent data for the 2020 Census. The Census Bureau places great value in the partnership and support provided by your networks, and we appreciate your collective commitment to helping the Census Bureau meet its dual mission of producing high quality statistics about the nation, while safeguarding the privacy of our respondents and the confidentiality of their data.

In your letter, you raised twenty-six questions about the Census Bureau's adoption of differential privacy and the implementation of the Disclosure Avoidance System. Enclosed, you will find our responses to your questions. I note that Census Bureau staff were available to discuss these and any related questions at the FSCPE Steering Committee meeting on February 12, 2020.

Sincerely,

John M. Abowd, PhD
Associate Director and Chief Scientist
Research and Methodology

Enclosure

Questions Regarding the Proposed Disclosure Avoidance System

a) To date, what was the process used for input in making the decision to implement DAS?

Answer: The U.S. Census Bureau's Data Stewardship Executive Policy Committee (DSEP) relies on input from a variety of sources when making decisions about the adoption, implementation, and parameters of the Disclosure Avoidance System (DAS). These include internal subject matter experts, the Census Bureau's advisory committees (the National Advisory Committee and the Scientific Advisory Committee), the Committee on National Statistics of the National Academy of Sciences, academic experts and researchers, privacy advocates, professional associations, federal and state partners, and many others. We also solicited public comment through a July 2018 Federal Register notice. The Census Bureau has also conducted formal tribal consultations with American Indian and Alaska Native tribal leaders. Engagement with these and other stakeholders is ongoing. The Census Bureau will continue to solicit and consider feedback to improve the DAS throughout the coming year.

b) What inputs (testimony, research, outside experts, etc.) did the Bureau's Data Stewardship Executive Policy Committee/Disclosure Review Board use to decide the final optimal privacy-loss budget (trade- off between privacy loss budget and data accuracy)?

Answer: The final privacy-loss budget for the 2020 Census has not yet been determined by the Data Stewardship Executive Policy Committee. The value of ε=6.0 was used for the release of the 2010 Demonstration Data Products, with ε=4.0 allocated to the person tables and ε=2.0 allocated to the household tables. These values were chosen after review of data presenting the impact on various demographic statistics computed for a range of different privacy-loss budgets.

c) Can you provide the planned scope of DAS project, the datasets and programs that will be affected by DAS (near and long-term), and the implementation schedule?

Answer: The DAS being developed for the 2020 Census of Population and Housing is being written specifically for the 2020 Census and cannot be directly applied to any other data product. However, the scientific and technical advances made during the development of the 2020 DAS will inform and enable the future development of formally private solutions for other Census Bureau products.

d) Does the Bureau have research that shows the 2010 Census file could be reconstructed / individuals identified without use of outside data files? It would be helpful for us to know if the reconstruction could be done without the use of a commercial dataset(s); and, if not, what commercial datasets the Census Bureau has used in reconstruction exercises.

Answer: The Census Bureau has now performed two partial reconstructions of the 2010 Census publications. Each reconstruction produces a microdata detail file containing the sex,

age, race, and ethnicity variables for each of the 308,745,538 enumerated individuals that made up the 2010 Census. No external data is required for the reconstruction attack. It uses only tract and block-level tables from the PL94-171 redistricting tables and the 2010 Summary File 1.

The reconstructed micro-data, which contain block, sex, age (in years) for every person enumerated in the 2010 Census, can then be matched against any file also containing block, sex and age to acquire names, and addresses for these same individuals. The attacker learns the exact race and ethnicity that was collected or imputed as part of the 2010 Census for specific individuals. The Census Bureau's internal re-identification experiments used a large database of commercial information that was acquired in the course of conducting the 2010 Census (the providers' names cannot be released because of the terms of the acquisition contract).

e) We understand that the Bureau is considering other formal privacy systems for public data products that include tables for detailed race and Hispanic origin tables, family/household tables that were included in Summary Files 1 and 2 in 2010, an American Indian and Alaska Native Summary File, and the Public Use Microdata Sample (PUMS) File. What are the other formal privacy systems under consideration for these data products? When will information about these systems and the availability of products be released and what will be the opportunity for input by the networks?

Answer: There are many different ways of implementing formal privacy, and the optimal selection and design of these systems depends on the characteristics of the data and the specific use cases for which data accuracy is to be optimized. The TopDown Algorithm (TDA) central to the DAS was designed specifically to produce the PL94-171 (redistricting data), the Demographic Profiles, and the demographic and housing characteristics data files. For other 2020 Census data products, including tables for detailed race and Hispanic origin, and tables previously included in the American Indian and Alaska Native (AIAN) Summary File, and tables requiring person/household joins, we are planning a secondary system that will be added as an extension of the DAS but still based on differential privacy. Because of their level of detail, these tables pose difficult and unique privacy challenges. The Census Bureau remains committed to producing data on detailed race and ethnicity, AIAN tribal affiliations, and characteristics of people within households to meet our data users' needs. Contracts have been executed to implement differentially private methods to produce these tables and to modify the dissemination system so that they can be released, once produced. We anticipate being able to provide more information about these efforts in the coming months, at which point we will work with the data user community to ensure these products meet their needs. At the present time, no final decisions have been made regarding the release of a Public Use Microdata Sample (PUMS) file for the 2020 Census.

f) How is the Bureau coordinating the implementation of DAS across divisions and branches? What are the contingency plans if DAS cannot be implemented as currently envisioned?

Answer: The DAS is being implemented by an agency-wide, interdisciplinary team that is headed by the Chief Scientist.

The DAS is currently operational and able to perform the disclosure avoidance necessary for the legislatively mandated publications of the decennial census. As we work to improve and optimize the DAS for an array of priority data use cases, we are also researching a variety of contingency plans to ensure that the 2020 Census Data Products meet the Census Bureau's data quality standards.

g) Can you provide specific information regarding the criteria that will determine suppression by geographic level? We are concerned that data may not available for key geographies including but not limited to blocks, block groups, tracts, census designated places, and minor civil divisions. It is of grave concern that DAS would have a substantial adverse impact on the availability and quality of data for small communities.

Answer: Suppression is a traditional disclosure avoidance technique that protects privacy by redacting or not publishing data for small groups or small geographies. One of the advantages of differential privacy is that noise infusion and the privacy guarantee removes the need for suppression by geographic level. Because of the impact of differential privacy on data accuracy for small geographies or populations, however, the Census Bureau is evaluating what tables to release and at what geographic levels to ensure that our data products meet fitness-for-use standards.

h) Can you provide similar information for other programs (American Community Survey (ACS), Economic Census, National Center for Health Statistics, etc.) on the tables or products that will be either modified or suppressed by using DAS or similar techniques, and the impact on geographies by FIPS and by NAICS for business data?

Answer: The application of differential privacy to the ACS, the Economic Census, and other data products is still in its research phase.

i) Will DAS also be implemented on data products produced for other agencies? Have these other agencies (such as BEA, the Department of Housing and Urban Development, the Bureau of Labor Statistics, the National Center for Education Statistics, and the National Science Foundation) and their data users provided comments? If so, in what form and where can the comments be accessed?

Answer: The 2020 DAS is custom developed for the 2020 Census and cannot be applied directly to multi-stage probability sample surveys, the bulk of the agency's contractual and internal products. More generally, the Census Bureau has indicated that it is committed to moving to formal privacy protection techniques for all of its data products, but the timetable for that

transition is unclear at present. We are committed to a vigorous and open discussion of the new DAS, and its impact on data quality and fitness-for-use. It has always been important for data users to understand the consequences of disclosure limitation procedures, and we know that these conversations will benefit the Census Bureau and improve our data products.

j) Will DAS be implemented on data products produced as a special tabulation? Has there been any research by the Bureau on special tabulation uses and the impact DAS has on these tabulations?

Answer: The mechanism by which formal privacy will be applied to special tabulations based on the 2020 Census is currently under development.

k) What criteria were used to decide on the proposed suite of the demographic and housing characteristic file (DHC) data tables?

Answer: To create the proposed suite of the demographic and housing characteristic file (DHC) tables, the Census Bureau started examining the suite of demographic and housing characteristics tables that were present in the 2010 Census SF1 and other data products and then began removing statistics that did not have a well-defined use case or that could not be readily computed using the 2020 DAS TDA. As we discuss in our response to question (e), we are developing alternate formally private systems to produce data products that cannot be computed using the TDA. Decisions have been made based on demonstrated use cases received through the July 2018 Federal Register Notice and through extensive stakeholder engagement and outreach by Census Bureau subject matter experts. If tables at particular geographic levels did not have a demonstrated use case, they were considered for removal from the DHC. This stakeholder outreach and engagement is ongoing, and we will continue to revise the proposed suite of data products as we receive additional information from our data users.

l) Has the Census Bureau considered expanding the privacy budget as the 2020 Census data becomes less current and therefore less valuable to potential reconstructors? Typically the Census Bureau does not issue Summary File 2 data until 2 or more years after the census is taken. Is there consideration in balancing privacy, currency, and accuracy in the implementation of the DAS?

Answer: The Census Bureau is currently engaging with a variety of stakeholders on this and related questions. No decisions have been made at this time.

m) For the tables that will be released for which DAS has been applied, what impact will it have on the ability to compare tables over time?

Answer: We recognize that the adoption of differential privacy for the 2020 Census will have implications on time-series and trend analysis when using the 2020 data products in

conjunction with data from prior censuses and surveys, and we will work with our data users to provide guidance on how to compensate for these effects. Each table in the DHC will be released a single time in 2021. It is premature to discuss the comparison of the 2020 DHC with either future versions of the ACS or with the tables that will be released for the 2030 Census.

n) Has the Bureau researched the impact of not producing data that was previously available? Has the Bureau considered that federal, state, and local programs in many cases, have statutory or administrative mandates requiring the use of Census data for funding and for reporting? Many of these requirements have been developed because accurate Census data provides a way to equitably distribute state and local revenues and services.

Answer: The Census Bureau is committed to continuing to produce the high quality data on which our many stakeholders have come to rely, consistent with our statutory obligations to protect the privacy of our respondents and the confidentiality of their data. To ensure that we are meeting our data users' needs, we have already engaged in extensive outreach to the data user community to catalog the various statutory and administrative uses of decennial census data. On July 19, 2018, the Census Bureau solicited "feedback from users on 2020 Data Products" (83 FR 34111). A second solicitation (83 FR 50636) appeared on October 9, 2018. The Census Bureau also participated in a meeting of the Committee for National Statistics on December 11-12, 2019, where major data users presented the results of their analyses of the 2010 Demonstration Data Products. More generally, the Census Bureau is eager to engage with federal, state and local programs to learn more of how they use census data and their requirements for accuracy. The Census Bureau is also eager to engage with stakeholders to understand the privacy expectations, requirements, and concerns of the American public. The Census Bureau's ultimate objective in engaging with these stakeholder groups is the development of principled approaches for balancing the needs and statutory requirements of both communities.

o) Most Census survey data already have margins of error (MOE). Does applying DAS compound these errors? The three networks already hear many complaints about MOEs – particularly for small geographies. We are concerned that, not only will the data be less usable, but that Census survey respondents will be unwilling to fill out forms if they perceive the resulting data is less accurate and unhelpful for their needs.

Answer: Error in census and survey statistics comes from a variety of sources (coverage error, measurement error, etc.), and privacy protections of all types add to this overall error. One major advantage of differential privacy over traditional disclosure avoidance techniques is that it allows the error (uncertainty) resulting from disclosure avoidance to be measured, reported, and discussed in a fully transparent manner. In general, the Census Bureau expects that the impact of the error introduced by the use of formal privacy will be less than the error resulting from other factors.

p) Why did the Bureau take the proactive stance to be the global leader in implementing disclosure avoidance, when so many vendors are collecting, selling, and publishing data that are not under the Bureau's control? It is important that our network members, data users, and the stakeholders we serve understand why the Bureau took the action proactively to be the global leader in disclosure avoidance without, as far as we know, any major challenge that the Bureau was not upholding its 13 USC mandate. It is also important to acknowledge that the Bureau's initiative will not solve the global problem of personal data disclosure.

Answer: The Census Bureau has been a global leader in the design and implementation of disclosure avoidance methods for decades. The Census Bureau's decision to adopt differential privacy is merely the next step in a history of innovation of our privacy protection methods to counter evolving privacy threats. The fact that there are many vendors collecting, selling and publishing data about United States residents does not lessen the Census Bureau's obligations under Title 13. Furthermore, although the coverage of commercial vendors is very good for some segments of the population, it is quite poor for others. Specifically, commercial data do not have good coverage of children, of self-reported race, of the existence of same-sex relationships, or of parents who have a different race than their children. This is precisely the sort of information that is collected by the decennial census. Given that the decennial census is a mandatory survey with universal coverage, the Census Bureau believes that it has both a legal and an ethical responsibility to use the strongest privacy protection technology available. 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, 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.

q) What other methods or consequences did the Bureau consider for protecting privacy, either legal or methodological, which would fulfill the Bureau's duty to protect an individual's record(s) and still produce data that can be used by the everyday user and local elected officials? How is this being handled by other governmental statistical agencies both within and outside of the US?

Answer: We know of no other statistical technique that can be reliably employed to assure the confidentiality of the underlying data while simultaneously assuring the highest quality statistical product for our data users. Other privacy mechanisms, such as the k-anonymity technique or even the swapping technique that was used in the 2010 Census, are now generally recognized by privacy researchers as being insufficient to meet twenty-first century privacy threats. In this era of Big Data, simply adding more noise using our older methods is not a workable solution. So much noise would be required that our published data would be unfit for most uses.

r) Has the Bureau considered the consequences of the implementation of DAS on non-governmental entities and programs that provide key community services? Has the Bureau presented the impacts and received input from small states, local, and non-profits on the implementation of differential privacy? A major concern is whether these organizations will have usable data to conduct research, present a case for grant funding, and build the right sized and type programs needed for their communities.

Answer: The Census Bureau is committed to publishing accurate data for the 2020 Census, however our obligations to protect privacy mean that we cannot publish perfectly accurate data for every conceivable use case. Based on the stakeholder feedback we have, and continue to, receive, we are endeavoring to ensure that the 2020 data products meet as many of our data users' needs as possible. The Census Bureau is also committed to maintaining the scientific integrity of the analyses performed using the public-use products that the Census Bureau releases. The Census Bureau will issue suitability for use guidelines that reflect the effects of the DAS. We will also publish our final algorithms, and the parameter values used by those algorithms, so that researchers can use the data in a scientifically appropriate manner. To conduct scientific analyses for which the public-use data are not suitable or sufficiently accurate, researchers may choose to seek approval to conduct a project under the auspices of the Federal Statistical Research Data Centers.

s) Has the Bureau determined the impact on program reports whose findings may be distorted due to the implementation of DAS and may no longer accurately represent the reported geographic area, population group, or economic sector?

Answer: Such work is currently ongoing. The Census Bureau is eager to work with stakeholders to develop systems such that the impact of the DAS on geographic areas, population groups, and economic sectors can be quantified and minimized.

t) Has the Bureau considered that the implementation of DAS will result in limited data availability for small geographies, leading these entities or service providers to purchase data or conduct surveys through private companies? The profusion of companies willing to provide data and surveys may compound disclosure issues since they are not subject to 13 USC requirements and will not use the same strict methods and guidelines the Bureau employs for both data collection and tabulation. For example, with the lack of state level population projections the private sector has stepped in from a variety of vendors with different products and levels of transparency.

Answer: Although the data we produce for the 2020 Census will be infused with noise, the 2020 DAS is designed such that statistics computed on larger populations, such as block groups or census tracts, will be significantly more accurate than statistics computed at the level of a single block. We agree that the proliferation of third-party data sources poses serious privacy concerns for our respondents. Since the last decennial census, the data world has changed dramatically. Growth in computing power, advances in mathematics, and easy access to large, public databases pose a significant threat to confidentiality. These forces have made it possible

for sophisticated users to ferret out common data points between databases using only our published statistics. If left unchecked, those users might be able to stitch together these common threads to identify the people behind the statistics. Because we are sworn by law to protect our respondents' data, we are constantly testing and improving our privacy protection methods to stay ahead of these changes. Our adoption of differential privacy for the 2020 Census is necessary to ensure that as more of these third-party data sources emerge and improve over time, they will not weaken or erode the privacy guarantees we provide to our respondents.

u) What plan does the Bureau have to inform and work with data users to ensure the implementation and impacts of DAS do not have negative consequences?

Answer: The Census Bureau is actively working to better inform and engage with our data users and the broader American public regarding all of the Census Bureau's efforts to protect respondent privacy while providing high-quality statistics about the nation. These ongoing efforts have taken many forms, including the December 2019 workshop sponsored by the National Academies' Committee on National Statistics. While many data users might wish for data to be published as accurately as possible—that is, without any privacy protection—there are also many respondents for whom privacy is a major concern, making the Census Bureau's ability to safeguard respondent data vital to the Census Bureau's efforts to maximize response rates. In the end, all statistical projects and efforts to protect respondent privacy have both positive and negative consequences that must be balanced by policy makers.

v) After data release, how is the Census Bureau going to handle criticism from the public that starts questioning the quality of the data because they find implausible numbers or don't recognize themselves or the area they live in in the published numbers? Will our networks get any guidance on how to deal with that criticism?

Answer: The Census Bureau is actively working to improve the 2020 DAS so that there will be few such implausible numbers in the 2020 data products.

w) What is the process, format, and timeline for the three networks to provide input to the Bureau? The input would include both the specific impact of DAS on the data for governmental and non- governmental organizations, as well as the result of network member comparative analyses of the demonstration tables and the 2010 tables.

Answer: Improvements and optimization of the DAS are ongoing, and will continue throughout most of calendar year 2020. Consequently, we welcome any data user feedback on the 2010 Demonstration Data Products that we may receive through the summer of 2020. While feedback received earlier in this process will have the greatest potential for informing major changes to the DAS design and configuration, we will assess and consider all feedback that we

receive on the Demonstration Products on an ongoing basis until the configuration of the DAS is finalized in late 2020. Partners may submit their feedback on the Demonstration Products individually or collectively by submitting them to dcmd.2010.demonstration.data.products@census.gov.

x) The Census Bureau has asked each of the partnerships to provide support for the implementation of differential privacy. Can you please provide what you are requesting each partnership to do to show support?

Answer: From our perspective, the most helpful feedback we could receive from the partnerships, in addition to the obvious identification of impossible or improbable outcomes in the 2010 Demonstration Data Products, would be suggestions that could be used to improve the design and optimization of the DAS to produce data products with the highest fitness-for-use.

With the understanding that there are basic tradeoffs between accuracy and privacy that DSEP will need to navigate, the most actionable suggestions we could receive from the partnerships would include results-oriented objectives (e.g., "willingness to sacrifice some existing accuracy at the block level to improve tract-level data") or standards-based thresholds (e.g., "county/tract/block-level data needs to be at least X/Y/Z percent accurate to be acceptable").

y) How does injecting noise into the data, disconnecting household relationship and effectively changing population counts for small areas impact the Census Bureau's residence rules and how local and state governments review the accuracy of the Census?

Answer: In past censuses, the Count Question Resolution (CQR) program has provided jurisdictions with the opportunity to verify the correct geolocation of group quarters facilities and housing units in Census Bureau tabulations. While the Census Bureau has not yet finalized details of how the CQR program will operate for the 2020 Census, we recognize that the operation of this program may be impacted by the transition to differential privacy. As these details are finalized we will engage with the partnerships to better answer this question.

z) How does the Bureau justify shrinking the availability of data about communities ranging from Asian ethnic groups and the Middle Eastern community when they have been asking for an expansion of how their specific communities are reported at least at a national level?

Answer: The Census Bureau is committed to producing data on detailed race, ethnicity, and American Indian and Alaska Native tribal affiliation at various geographic levels to meet our data users' needs. As these products will be produced through a different formally private system separate from the DAS, they are not currently included in the 2010 Demonstration Data Products. As we continue with the design and development of this second system over the coming months, we will actively engage with the partnerships to evaluate and improve these additional data products.

Correspondence Quality Assurance Staff

Office of the Director
U.S. Census Bureau

Control Sheet

November 27, 2019

Census Id: CQAS-09930

DOC Number:

Correspondence Type: Controlled Correspondence

Action Office: ADRM

Signature: ~~Dillingham~~ Abowd

Subject: Census Bureau's key partnership programs (Census Information Centers (CIC), Federal-State Cooperative for Population Estimates (FSCPE), and State Data Centers (SDC)) present this letter outlining our questions and concerns regarding the Census Bureau's move to a new Differential Privacy Disclosure Avoidance System (DAS)

Instructions: Prepare draft and send to CQAS.

Due in CQAS: 12/10/2019

Sender: CIC Steering Committee, FSCPE Steering Committee, SDC Steering Committee

Constituent:

Corr Date: 11/27/2019

Rec Date: 11/27/2019

Due Date: 12/16/2019

Confidential Information: No

Addressee: Dillingham

Infocopy: CLMSO,BUCKNER,Stanley,Tadlock

12/30 Rejected. Sent comment to ADRM
2/3: Sent A/o new draft to LegaCads
Nick will look at it after Legal's comment addressed. Conn 12/31
12/23

Concurrence

Surname	CED	CQAS	Tadhawk	Legal	DCO			Abowd
Initials	Ms Hawes	AJCW	NTH	MW				
Date	12/20	12/23	12/31	2/3				

Fwd: Joint Letter Regarding DAS

Christopher J Stanley (CENSUS/OCIA FED)
Wed 11/27/2019 12:42 PM
To: BOC Correspondence Quality Assurance (CENSUS) <boc.correspondence.quality.assurance@census.gov>

2 attachments (166 KB)
2019_11_27-Joint Letter to the Director.pdf; ATT00001.htm;

Please control this one to R&M as the action office. Please copy Buckner and Misty Reed on the assignment. Thank you very much.

Begin forwarded message:

From: "Stephen L Buckner (CENSUS/ADCOM FED)" <Stephen.L.Buckner@census.gov>
Date: November 27, 2019 at 12:34:09 PM EST
To: ADCOM Issues Working Group List <adcom.issues.working.group.list@census.gov>
Subject: Fw: Joint Letter Regarding DAS

SDC/CIC/FSCPE letter to the director about DP.

Stephen

Stephen L. Buckner, Assistant Director
Communications Directorate
U.S. Census Bureau
O: 301-763-3586 | M: 301-792-6587
census.gov | @uscensusbureau

From: Schenker, Pamela <SCHENKER.PAMELA@leg.state.fl.us>
Sent: Wednesday, November 27, 2019 12:26 PM
To: Steven Dillingham (CENSUS/DEPDIR FED) <steven.dillingham@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Stephen L Buckner (CENSUS/ADCOM FED) <Stephen.L.Buckner@census.gov>; Misty L Reed (CENSUS/CLMSO FED) <Misty.L.Reed@census.gov>; Lakiva M Pullins (CENSUS/CLMSO FED) <Lakiva.M.Pullins@census.gov>; Leland Todd Webb (CENSUS/PPSI FED) <Leland.Todd.Webb@census.gov>; Karen Battle (CENSUS/POP FED) <karen.battle@census.gov>; Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>; Marc J Perry (CENSUS/POP FED) <Marc.J.Perry@census.gov>; Rachel Marks (CENSUS/POP FED)

<Rachel.Marks@census.gov>; Kevin Barragan (CENSUS/POP FED) <kevin.barragan@census.gov>; jhardcastle@tax.state.nv.us <jhardcastle@tax.state.nv.us>; elizabeth.garner@state.co.us (CENSUS/ OTHER) <elizabeth.garner@state.co.us>; jkv3@cornell.edu <jkv3@cornell.edu>; sstrate@donahue.umassp.edu <sstrate@donahue.umassp.edu>; Jim.Chang@oeo.az.gov <Jim.Chang@oeo.az.gov>; rhatigan@unm.edu <rhatigan@unm.edu>; guthriee@michigan.gov (CENSUS/ OTHER) <guthriee@michigan.gov>; Census CIC SteeringCommittee List <census.cic.steeringcommittee.list@census.gov>; howard.shih@aafederation.org (CENSUS/ OTHER) <howard.shih@aafederation.org>; Allen <Allen.Barnes@oeo.az.gov>; bob.coats@osbm.nc.gov (CENSUS/ OTHER) <bob.coats@osbm.nc.gov>; jjb131@psu.edu (CENSUS/ OTHER) <jjb131@psu.edu>; Mallory.bateman@utah.edu (CENSUS/ OTHER) <Mallory.bateman@utah.edu>; Mary.Craigle@mt.gov <Mary.Craigle@mt.gov>; mmoser@uvm.edu (CENSUS/ OTHER) <mmoser@uvm.edu>; sreagan@unm.edu (CENSUS/ OTHER) <sreagan@unm.edu>; todd.graham@metc.state.mn.us (CENSUS/ OTHER) <todd.graham@metc.state.mn.us>
Subject: Joint Letter Regarding DAS

Director Dillingham –

The steering committees of the Census Information Centers (CIC), Federal-State Cooperative for Population Estimates (FSCPE), and State Data Centers (SDC)) present this attached letter outlining our questions and concerns regarding the Census Bureau's move to a new Differential Privacy Disclosure Avoidance System (DAS). We look forward to your responses and opportunities to discuss the issues with you and your staff.

Our steering committees can be reached by email: jhardcastle@tax.state.nv.us, or by US mail: Jeff Hardcastle, State Demographer, 4600 Kietzke Lane, Building L, Suite 235, Reno, NV 89502.

Thank you, and as your partners, we strive to give maximum support all of the Census Bureau's work and the success of the 2020 Census.

THE STEERING COMMITTEES OF

CENSUS INFORMATION CENTERS

FEDERAL STATE COOPERATVE FOR POPULATION ESTIMATES

STATE DATA CENTERS

November 27, 2019

Steven Dillingham, Director
U.S. Census Bureau
4600 Silver Hill Road, Room 8H001
Washington, DC 20233

Dear Director Dillingham,

The members of the Census Bureau's key partnership programs (Census Information Centers (CIC), Federal-State Cooperative for Population Estimates (FSCPE), and State Data Centers (SDC)) present this letter outlining our questions and concerns regarding the Census Bureau's move to a new Differential Privacy Disclosure Avoidance System (DAS). As the Bureau's premier partners and supporters, the three networks want to ensure that we fully understand the Bureau's decision process. We have concerns that this implementation has been driven by data scientists with limited consideration for users' needs. We are particularly concerned that insufficient analysis has been conducted regarding how DAS will affect the Census data used for informing policy and allocating public and private funds. We hope to broaden the discussion and raise awareness of the impacts of DAS on state and local decision-making.

Our network members understand the Bureau's objective to balance privacy with data usability and availability. However, the repercussions and loss of data should be carefully weighed and evaluated relative to the Bureau's responsibilities for privacy protection under 13 USC § 9. Our committees and networks have concerns and questions regarding the timing, breadth, and scope of DAS as implemented by the Bureau. Further, we have concerns that the proposed implementation violates the Census Bureau's obligation, under 13 USC § 141, to provide a Redistricting Data File with accurate population counts. It is only recently that we have had enough specifics to analyze the potential impacts and provide meaningful feedback to the Bureau. The data released however is not complete as many issues are not yet resolved.

Our three partnerships are fully engaged in public outreach and promotion of the 2020 Census. Our communications have emphasized the goal of reliable public data. If DAS impacts the numbers to the point that the data is not fit for use, or not available for small communities, then the Census Bureau's standing as the gold standard for data will be diminished.

We are asking that the Bureau provide a detailed response and time for discussion of the points we raise in this letter. We would like to have this discussion before year-end. There is considerable concern about Differential Privacy DAS, its origins as a policy, and its implementation. This extensive change will only be successful with full vetting across the user communities and the Bureau's other federal partners. We hope that this discussion with the Bureau can produce a policy direction and DAS implementation that serves public data users before there are adverse effects.

Thank you for your attention to our questions and comments. As the Census Bureau's key partners, major users and disseminators of census data, we look forward to your responses and discussions. Our steering committees can be reached by email: jhardcastle@tax.state.nv.us, or by US mail: Jeff Hardcastle, State Demographer, 4600 Kietzke Lane, Building L, Suite 235, Reno, NV 89502. Our network members are here to support the Census Bureau's work and the success of the 2020 Census.

Sincerely,

CIC Steering Committee
FSCPE Steering Committee
SDC Steering Committee

cc: John Abowd, Associate Director for Research and Methods
Stephen L. Buckner, Assistant Director of Communications
Misty L. Reed, Division Chief, Customer Liaison & Marketing Services Office
Lakiva N. Pullins, Assistant Division Chief, Customer Liaison and Marketing Services Office
Leland Todd Webb, Chief, Data Users Branch
Karen Battle, Chief, Population Division
Michael B. Hawes, Senior Advisor for Data Access and Privacy, Research and Methodology Division
Marc J. Perry, Senior Demographic Reviewer, Population Division
Rachel Marks, Senior Technical Expert on Population Statistics, Population Division
Kevin Barragan, Statistician/Demographer, Coordination, Dissemination, and Outreach Branch, Population Division
CIC Steering Committee and Members
FSCPE Steering Committee and Members
SDC Steering Committee and Members

Attachment: (3 pages)

Questions Regarding the Proposed Disclosure Avoidance System (DAS)

a) To date, what was the process used for input in making the decision to implement DAS?

b) What inputs (testimony, research, outside experts, etc.) did the Bureau's Data Stewardship Executive Policy Committee/Disclosure Review Board use to decide the final optimal privacy-loss budget (trade-off between privacy loss budget and data accuracy)?

c) Can you provide the planned scope of DAS project, the datasets and programs that will be affected by DAS (near and long-term), and the implementation schedule?

*d) Does the Bureau have research that shows the 2010 Census file could be reconstructed / individuals identified **without** use of outside data files?*
It would be helpful for us to know if the reconstruction could be done without the use of a commercial dataset(s); and, if not, what commercial datasets the Census Bureau has used in reconstruction exercises.

e) We understand that the Bureau is considering other formal privacy systems for public data products that include tables for detailed race and Hispanic origin tables, family/household tables that were included in Summary Files 1 and 2 in 2010, an American Indian and Alaska Native Summary File, and the Public Use Microdata Sample (PUMS) File. What are the other formal privacy systems under consideration for these data products? When will information about these systems and the availability of products be released and what will be the opportunity for input by the networks?

f) How is the Bureau coordinating the implementation of DAS across divisions and branches? What are the contingency plans if DAS cannot be implemented as currently envisioned?

g) Can you provide specific information regarding the criteria that will determine suppression by geographic level?
We are concerned that data may not available for key geographies including but not limited to blocks, block groups, tracts, census designated places, and minor civil divisions. It is of grave concern that DAS would have a substantial adverse impact on the availability and quality of data for small communities.

h) Can you provide similar information for other programs (American Community Survey, Economic Census, National Center for Health Statistics, etc.) on the tables or products that will be either modified or suppressed by using DAS or similar techniques, and the impact on geographies by FIPS and by NAICS for business data?

i) Will DAS also be implemented on data products produced for other agencies? Have these other agencies (such as BEA, the Department of Housing and Urban Development, the Bureau of Labor Statistics, the National Center for Education Statistics, and the National Science Foundation) and their data users provided comments? If so, in what form and where can the comments be accessed?

j) Will DAS be implemented on data products produced as a special tabulation? Has there been any research by the Bureau on special tabulation uses and the impact DAS has on these tabulations?

k) What criteria were used to decide on the proposed suite of the demographic and housing characteristic file (DHC) data tables?

l) *Has the Census Bureau considered expanding the privacy budget as the 2020 Census data becomes less current and therefore less valuable to potential reconstructors? Typically the Census Bureau does not issue Summary File 2 data until 2 or more years after the census is taken. Is there consideration in balancing privacy, currency, and accuracy in the implementation of the DAS?*

m) *For the tables that will be released for which DAS has been applied, what impact will it have on the ability to compare tables over time?*
The networks have already identified unacceptable inconsistencies when comparing the demonstration products with previously published 2010 data.

n) *Has the Bureau researched the impact of not producing data that was previously available? Has the Bureau considered that federal, state, and local programs in many cases, have statutory or administrative mandates requiring the use of Census data for funding and for reporting?*
Many of these requirements have been developed because accurate Census data provides a way to equitably distribute state and local revenues and services.

o) *Most Census survey data already have margins of error (MOE). Does applying DAS compound these errors?*
The three networks already hear many complaints about MOEs – particularly for small geographies. We are concerned that, not only will the data be less usable, but that Census survey respondents will be unwilling to fill out forms if they perceive the resulting data is less accurate and unhelpful for their needs.

p) *Why did the Bureau take the proactive stance to be the global leader in implementing disclosure avoidance, when so many vendors are collecting, selling, and publishing data that are not under the Bureau's control?*
It is important that our network members, data users, and the stakeholders we serve understand why the Bureau took the action proactively to be the global leader in disclosure avoidance without, as far as we know, any major challenge that the Bureau was not upholding its 13 USC mandate. It is also important to acknowledge that the Bureau's initiative will not solve the global problem of personal data disclosure.

q) *What other methods or consequences did the Bureau consider for protecting privacy, either legal or methodological, which would fulfill the Bureau's duty to protect an individual's record(s) and still produce data that can be used by the everyday user and local elected officials? How is this being handled by other governmental statistical agencies both within and outside of the US?*

r) *Has the Bureau considered the consequences of the implementation of DAS on non-governmental entities and programs that provide key community services? Has the Bureau presented the impacts and received input from small states, local, and non-profits on the implementation of differential privacy?*
A major concern is whether these organizations will have usable data to conduct research, present a case for grant funding, and build the right sized and type programs needed for their communities.

s) *Has the Bureau determined the impact on program reports whose findings may be distorted due to the implementation of DAS and may no longer accurately represent the reported geographic area, population group, or economic sector?*

t) *Has the Bureau considered that the implementation of DAS will result in limited data availability for small geographies, leading these entities or service providers to purchase data or conduct surveys through private companies?*
The profusion of companies willing to provide data and surveys may compound disclosure issues since they are not subject to 13 USC requirements and will not use the same strict methods and guidelines the Bureau employs for both data collection and tabulation. For example, with the lack of state level population

projections the private sector has stepped in from a variety of vendors with different products and levels of transparency.

u) *What plan does the Bureau have to inform and work with data users to ensure the implementation and impacts of DAS do not have negative consequences?*

v) *After data release, how is the Census Bureau going to handle criticism from the public that starts questioning the quality of the data because they find implausible numbers or don't recognize themselves or the area they live in in the published numbers? Will our networks get any guidance on how to deal with that criticism?*

w) *What is the process, format, and timeline for the three networks to provide input to the Bureau?* The input would include both the specific impact of DAS on the data for governmental and non-governmental organizations, as well as the result of network member comparative analyses of the demonstration tables and the 2010 tables.

x) *The Census Bureau has asked each of the partnerships to provide support for the implementation of differential privacy. Can you please provide what you are requesting each partnership to do to show support?*

y) *How does injecting noise into the data, disconnecting household relationship and effectively changing population counts for small areas impact the Census Bureau's residence rules and how local and state governments review the accuracy of the Census?*

z) *How does the Bureau justify shrinking the availability of data about communities ranging from Asian ethnic groups and the Middle Eastern community when they have been asking for an expansion of how their specific communities are reported at least at a national level?*

EXHIBIT 21

POLIDATA® Political Data Analysis

DATABASE DEVELOPMENT, ANALYSIS AND PUBLICATION;
POLITICAL AND CENSUS DATA; LITIGATION SUPPORT

CLARK BENSEN

POLIDATA LLC · 1303 HAYWARD RD, P.O. BOX 530 · CORINTH, VT 05039
Tel: 703-690-4066 · Fax: 202-318-0793 · email: clark@polidata.org
PUBLISHER OF THE POLIDATA ® DEMOGRAPHIC AND POLITICAL GUIDES AND ATLASES

Honorable Steven Dillingham, Director
U. S. Bureau of the Census
4600 Silver Hill Road
Washington, DC 20233

10 Apr 2020

Re: DAP2020

Director Dillingham,

This letter raises some concerns that I, as one who has been involved in districting projects since the 1980 Census, has about the Disclosure Avoidance Program (DAP). This is briefly described on a Bureau webpage entitled, "Statistical Safeguards":

> *Before we publish any statistic, we apply safeguards that help prevent someone from being able to trace that statistic back to a specific respondent.*
>
> *We call these safeguards "disclosure avoidance," although these methods are also known as "statistical disclosure controls" or "statistical disclosure limitations."*
>
> *Although it might appear that a published table shows information about a specific individual, the Census Bureau has taken steps to disguise the original data in such a way that the results are still useful. These steps include using statistical methods such as "data swapping" and "noise injection."*

Before Census 2000 a similar issue faced the Bureau with regards to adjustment of the census counts. Congress enacted a statute[1] which addressed "Statistical Sampling or Adjustment" in the decennial. Important concerns of Congress expressed in the findings to PL105-119 are: *"(5) the decennial enumeration of the population is one of the most critical constitutional functions our Federal Government performs; (6) it is essential that the decennial enumeration of the population be as accurate as possible, consistent with the Constitution and laws of the United States[;]."*

The Supreme Court addressed that situation in an opinion announced on January 25, 1999[2], 14 months before Census Day 2000, *"States use the population numbers generated by the federal decennial census for federal congressional redistricting. See Karcher v. Daggett, 462 U. S. 725, 738 (1983) ("[B]ecause the census count represents the 'best population data available,' . . . it is the only basis for good-faith attempts to achieve population equality"…).*

While the Commerce case focused largely on sampling, the act is more expansive and another of its findings is: *"(7) the use of statistical sampling or statistical adjustment in conjunction with an actual enumeration to carry out the census with respect to any segment of the population poses the risk of an inaccurate, invalid, and unconstitutional census[;]."* A review of the language in section (h) of the findings provides a definition of what the term 'statistical method' means. This definition includes "or any other statistical

[1] See Pub. L. 105-119; Sec. 209 (a) (5) [congressional findings] Statistical sampling or adjustment in decennial enumeration of population; https://uscode.house.gov/statviewer.htm?volume=111&page=2480

[2] See *Department of Commerce v. United States House of Representatives*, 525 US 316 (1999); (98-404); argued November 30, 1998; decided January 25, 1999.

procedure, including statistical adjustment, to add or subtract counts to or from the enumeration of the population as a result of statistical inference[;]."

My main concern is with respect to districting[3] and is that if the Bureau implements the DAP as it is currently envisioned the thousands of entities across the nation that are responsible for revising current, or creating new district, boundaries for representative government at the state and local level will not have the "best population data available" and therefore will not be able to make good-faith attempts towards equality. I offer these comments with the understanding that many of the general concerns will be shared by numerous redistricting stakeholders once they know about DAP. Moreover, I believe there is general agreement regardless of political affiliation on this issue.

This is simply a question of process. The entities responsible for districting need to know, before the numbers are released in less than one year, that the numbers they receive will be sufficient to meet their critical need and that their own election calendars will not be disrupted by additional litigation over the numbers used to distribute political representation across their states or localities.

This is not a concern about the goals of the DAP to avoid inadvertent disclosure of personally identifiable information (PII). I believe there is substantial agreement that the privacy of certain individuals is a laudable aim in 2020[4]. However, it appears that the DAP presents a fundamental interference with the constitutional purposes of apportionment by reliance upon a statutory concern relating to privacy.

While a supplement to this letter will discuss some of the concerns shared by redistricting stakeholders, they will be listed below.

1) Adjusted numbers will not be "the best available population data".
2) Stakeholders will be unable to "make good faith efforts" at equality.
3) Use of such a statistical method "poses the risk of an inaccurate, invalid, and unconstitutional census".
4) Additional litigation over the numbers may result in distraction, delay, and costs to many districting entities.
5) The confidence amongst state and local governmental entities in the entire census process may be severely undermined.
6) While the Bureau is a national statistical agency, first and foremost it is the compiler of the "actual Enumeration" to fulfill the constitutional mandate.
7) Previous methods for disclosure avoidance were less pervasive. Because the previous methods were simpler techniques such as data swapping, rounding, top-coding, etc., the degree to which information was adjusted for protection was much less. On the other hand, the DAP for 2020 will affect every level of geography and the population counts.
8) Relative inaccuracy and bias in the DAP: "The new method allows us to precisely control the amount of uncertainty that we add according to privacy requirements."

As discussed above, the implementation of the DAP is quite likely to affect redistricting stakeholders across the nation. It appears that there are several options available to the Bureau at this point.

1) **Continue with research but still implement DAP.** Of course, the Bureau could discount the concerns of the (currently) small group of stakeholders and local statistical entities and

[3] However, given the feedback from the so-called Demonstration Data during 2019 there are other concerns, such as distribution of intergovernmental aid, that may motivate others to comment on the DAP.
[4] Nevertheless, privacy was not an issue when the census was first taken. In fact, the first Census Act required the schedules to be posted for public review before they were submitted to the federal marshal. Specific requirements for privacy appear to have first been codified for the 1880 Census.

proceed as currently planned. Nevertheless, based upon the most recent information from working groups it appears that while improvements may be made to the range of error introduced by noise injection, the counts will still not be available for most levels of geography.

2) **The Black Box Engine.** Some observers have suggested that districting entities could submit any plan of interest to a website whereby the unadjusted counts could be applied and thus the plan drafters could know expeditiously how far off their numbers were from equality. Aside from the obvious logistical issues for such a process it fails for the want of transparency.
3) **Reduce the cross tabulations of data tables.** This could apply in a general sense to whatever cross tabulations that the Bureau provides. Such breakdowns appear to be largely developed by the Bureau for the use of federal, state, and local governments in their mission to fulfill their requirements for purposes other than apportionment.
4) **Reduce the breakdowns of data tables into fewer cells.** The critical dataset for redistricting, the so-called PL94 dataset[5] was, prior to Census 2000, a fairly simple dataset with a much smaller set of variables. With the addition of the multi-race response options in 2000 the number of data cells for the PL dataset expanded greatly. On its face this presents numerous privacy concerns even for areas that have a substantial number of persons because all six races are tallied for all multiple combinations. The level of detail in the PL94 dataset for each record is not needed by most districting entities and could be collapsed substantially and then DAP adjustments as previously done to the characteristic data could be undertaken.
5) **Invariant Block Counts without Characteristic Information.** Another alternative would be to hold invariant the counts of population and housing[6] and to simply provide no characteristic information at the block level. Choices for such an alternative could be a) include characteristic data only for areas at a specified geographic level or with counts above a threshold, as has been done with Special Tabulations previously, and/or b) have districting entities rely upon characteristic data from the American Community Survey (ACS).

Clearly, the perspective of districting stakeholders and local planning agencies is likely to something other than Option 1[7]. Because districting is done for so many types of entities there are varying degrees of resources and needs. Yet, considering the range of variations that are likely to be seen when a user compares the adjusted numbers to information they have independently collected over the decade, there are going to be a lot of queries. One would expect that local officials may find significant differences because they can spend the time to review the information, block by block. What does the Bureau propose for the Count Question Resolution process for Census 2020?

Other stakeholders may weigh in on this issue as well offering different options or perspectives. However, Options 4 and 5 at least appear to several stakeholders as being viable options. Option 4 could impose a burden on a relatively small number of entities but may not appease the concerns of the Bureau for privacy. Option 5 would affect substantially more entities but at least there is some alternative source of data that would provide less precision for the characteristic data and more statistical analysis for districting entities to comply with Voting Rights Act concerns. Nevertheless, even accepting Options 4 or 5 would be a substantial compromise for some stakeholders but if the only viable option for privacy is the DAP many stakeholders would likely choose one of the above or some other alternative not yet discussed.

---

[5] See Pub. L. 94-171. https://uscode.house.gov/statutes/pl/94/171.pdf
[6] Total Population and Voting Age Population, as well as the information on Housing Units and Group Quarters.
[7] N.b., while there may not be much difference of opinion about the overall concern, there may well be with respect to options.

Respectfully yours,

/s/ Clark H. Bensen

Clark H. Bensen

Enclosures:
1) Supplement

[2020-0410a]

CC:
Honorable Wilbur Ross, Secretary
U.S. Department of Commerce
1401 Constitution Ave, NW
Washington, DC 20230

SUPPLEMENT

Introduction. For the sake of readers of this letter for whom Disclosure Avoidance is a new concept the following brief summary is provided. It is important to understand the widespread degree to which the counts from the 'actual Enumeration' are likely to be affected by the DAP.

In December of 2019 a conference was held that reviewed the results from the Bureau's efforts of the application of the DAS to the 2010 Census data. Based upon information published by the Bureau during October 2019[8] and additional material published subsequent to the December 2019 conference and recent meetings of the Expert Group (which now includes at least one for redistricting) it is still unclear exactly what the actual plan for the Bureau is or will be. Moreover, it appears that the current schedule is that final policy decisions will not be made, for the design of the DAS, until September 2020[9].

Currently, the best information of the degree to which numbers eventually reported for the 2020 Census can only be gleaned from the information provided in the October 2019 memo which detailed the status of these numbers for the review of the 2010 Census data. In other words, the plan, at that point, was that some numbers would be 'invariant', that is, the reported number would be the enumeration counts and no alteration for privacy would be made, while others will be 'variant', that is, the numbers reported would be altered for privacy protection.

That proposal would treat only three types of counts as invariant: a) the state total population; b) the number of housing units in a census block; and c) the number and type of group quarters in a census block[10]. In other words, below the state, every number provided by the Bureau will not be a tabulation of the responses from an 'actual Enumeration' but the result of a statistical alteration. "Differential privacy allows us to inject a precisely calibrated amount of noise into the data to control the privacy risk of any calculation or statistic."[11]

Additionally, there is the question as to which metrics will be released with the adjusted numbers to allow users to assess the degree to which noise has been added. A recent March 2020 presentation[12] primarily addressed "making population counts more accurate" and reviewed numerous metrics that might "allow the public to see the improvements that are made" as the Bureau continues to test their DAS operations.

At this point it is an open question as to whether this will substantially change so that the block counts would be delivered as enumerated or adjusted. Regardless, what this indicates is that we are now less than one year away from releases of the numbers and the Bureau still does not know with any precision what method they will use or metrics they will provide. Notably, the implementation of disclosure avoidance will not be applied to the American Community Survey (ACS) until 2025[13]. Why is it that the purposes of apportionment will be the first real test case for such a statistical adjustment?

---

[8] See Memorandum 2019.25: 2010 Demonstration Data Products – Design Parameters and Global Privacy-Loss Budget; https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_25.html

[9] See Updates and DAS Development Schedule, March 18, 2020; https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-updates-das-development-schedule.pdf?#

[10] See the Bureau site: https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

[11] See the Bureau site: https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf?#

[12] See 2020 Census Disclosure Avoidance Improvement Metrics; https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf?#

[13] See https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html

One issue that appears to have concerned the Bureau over the threat of what they term as reconstruction of the census appears to be the result of the extraordinary level of detail that is provided by two data products: a) the block-level data provided pursuant to PL94-171 and b) the numerous cross-tabulation tables that are provided by the Bureau of data at numerous levels of census geography.

The block-level data is the critical dataset for most redistricting stakeholders. Blocks have a huge range of population across any geographic area. Many have no population because they are industrial areas, or parks, or bodies of water, or highways, or mountains, or wide open range, or simply vacant housing. Many have a handful, and many have thousands, of persons. But, blocks are used as the lowest level for most districting datasets, generally because of a few factors that make them unusual amongst all the so-called 'summary levels' that the Bureau recognizes.

These characteristics of census blocks include:

1) they are the lowest level for which the counts have heretofore been tabulated and made available;
2) they cover the entire non-coastal geographic area of a state or locality;
3) pursuant to the Block Boundary Suggestion Project (BBSP) the states have the ability to designate the boundaries of the blocks;
4) tabulations generally account for how the block fits into higher levels of geography, such as Voting Districts (VTDs) the boundaries of which are designated by many states as Phase 2 of the BBSP;
5) the reported counts for every higher level of geography has been simply the sum of the information for all corresponding blocks;
6) redistricting stakeholders form one of the few groups that rely upon the block-level information as the critical data needed to fulfill their need, that is, the purposes of apportionment; equalizing population would be considerably more difficult if higher level information was the only level for which accurate data were available[14].

Below are some notes on the concerns enumerated in the letter.

1) Adjusted numbers will not be "the best available population data".
    a. This is the language used in the *Karcher* case which was quoted by the SCOTUS in the Commerce Department opinion in 1999 about adjustment.
    b. The basic concern here is that the both phases of the apportionment process, i.e., the apportionment of seats to predetermined units (e.g., states) and the districting phase should rely upon the best available data.
    c. The Bureau has indicated that the state-level counts would be held invariant; a position that changed after initial discussions with stakeholders.
2) Stakeholders will not be able to "make good faith efforts" at equality.
    a. This language also refers to the *Karcher* case which basically requires a zero-tolerance for population amongst congressional districts.
    b. Also of note are the *Larios v. Cox* case (out of Georgia) in 2004[15] and the *Tennant v. Jefferson County Commission* case (out of West Virginia) in 2012[16]. Larios reiterated the focus of the reapportionment cases of the 1960s that the goal (therein for legislative districting) was to have equally populated districts.

[14] Note also that blocks are numbered by the Bureau and thus Block Groups, the next higher level above Blocks, are simply agglomerations of adjacent Blocks for statistical purposes. Census Tracts, the next level up the main hierarchy (aka the Spine) are designed to be generally consistent over time but have, on average thousands of persons.
[15] See *Cox v. Larios*, 542 US 947 (2004); no. 03-1413, decided June 30, 2004;
[16] See *Tennant v. Jefferson County Commission*, 567 US 758 (2012); no. 11-1184; decided September 25, 2012.

   c. The West Virginia case muddied this up a bit (for congressional districting) allowing some leniency for population deviation based upon the competing interests of the lowest deviation and legitimate state objectives. In reality this opinion reminded stakeholders of the original perspective of the Court in *Karcher*.

3) Use of such a statistical method "poses the risk of an inaccurate, invalid, and unconstitutional census".
   a. In its findings, the Congress was apparently referring to the competing analyses of the proposed adjustment for undercount which adjustment was to be based upon a statistical method known as sampling.
   b. The Commerce case hinged largely on the statutory interpretation of the Census Act in sections 141 and 195 and held that the statistical method known as sampling was not an available method for the numbers compiled for the purposes of apportionment.
4) Additional litigation over the numbers may result in distraction, delay, and costs to many districting entities.
   a. National entities are frequently at the forefront of litigation over these types of issues and bear the cost of having the courts reach a generally applicable ruling. However, given the range of error that might be infused into the process by noise injection it is likely that numerous cases may occur because of a dispute over how to interpret the altered numbers. The burden and confusion in such cases may redound to localities that may not be able to afford litigation through the entire process.
5) The confidence amongst state and local governmental entities in the entire census process may be severely undermined.
   a. Local officials will review the census results block-by-block and when they discover that the reported results are different, and frequently substantially so, they will be concerned.
   b. In recent censuses there has been a Count Question Resolution Program (CQR) to review the counts upon request and correct them if and as needed. It is unclear how this can be implemented if DAP is used for 2020.
6) While the Bureau is a national statistical agency, first and foremost it is the compiler of the "actual Enumeration" to fulfill the constitutional mandate.
   a. There appears to be a break in the internal firewall at the Bureau vis-à-vis fulfillment of the constitutional mandate and ongoing survey programs. Admittedly, the number of survey programs that are done for other agencies and those that present the demographics of the nation to the world are the everyday projects for much of the Bureau. Understandably, what is good enough for a statistical agency to present may fall short of the standard of care for the counts used for "the purposes of apportionment".
   b. Of course, there are some projects that focus on the high quality of the actual enumeration at the Bureau and Complete Count Committees, as well as NGOs, work diligently throughout the decade to make the decennial "the best population data available". Implementing DAP may lessen that focus because the numbers that will be used for redistricting will not be from the enumeration but altered in the manner proposed by the data scientists and decided by the Disclosure Review Board.
7) Previous methods for disclosure avoidance were less pervasive.
   a. The previous methods were simpler, and easily understandable, techniques such as data swapping, rounding, top-coding, etc. and the degree to which all census

information was adjusted for protection was much less. On the other hand, the DAP for 2020 will affect almost every level of geography and the population counts.

b. The DAP really is a 'sea change' for redistricting and the census. Users of the special tabulations have accepted previous efforts at disclosure avoidance because those users are cognizant of the problems and the shortcomings in protected data for their specific purpose, which would rarely require the precision needed for the purposes of apportionment.

8) Relative inaccuracy and bias in the DAP.

a. "The new method allows us to precisely control the amount of uncertainty that we add according to privacy requirements." Not only will the data scientists determine the best method to adjust the counts but there will inevitably be some loss of accuracy which will have some level of bias for or against some subgroup of the census universe.

b. It is still unclear exactly what this bias will be at this point but what is likely is that once a bias is anticipated or observed the question of using the DAP will no longer be simply one of process but a political fight of the disfavored groups against the favored groups.

###

[2020-0410a]

EXHIBIT 22

Formal Privacy At Scale: Reducing the Magnitude of Upward Bias

Philip Leclerc
On Behalf of & With the support of the 2020 Decennial Census Disclosure Avoidance System (DAS) development team
Center for Enterprise Dissemination-Disclosure Avoidance
U.S. Census Bureau
Symposium on Data Science & Statistics
May 22, 2020

The views in this presentation are those of the author, and not those of the U.S. Census Bureau.

Census TopDown Algorithm (TDA): An Overview of Its Structure & Properties

Census TDA: Requirements and Properties I

TDA is the principal formally private 2020 Census disclosure limitation algorithm under development

Inputs:

- Post-edits-and-imputation microdata records (Census Edited File – CEF)
- Required structural zeros & data-dependent invariants

Processing:

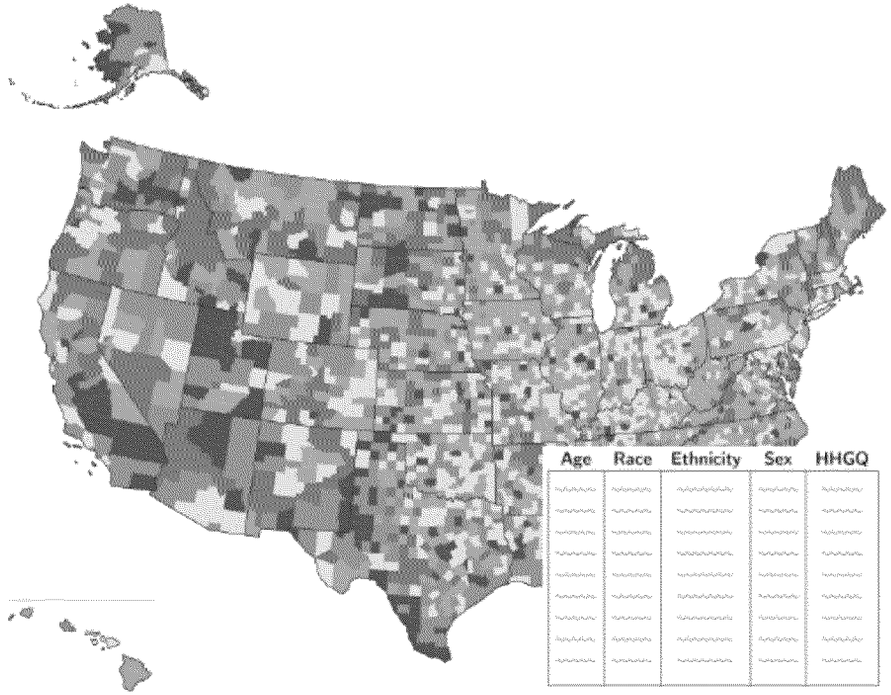
- Convert CEF to an equivalent "histogram" (fully saturated contingency table)
- Apply DP measurements & perform mathematical optimization
- Create noisy histogram; convert back to microdata

Output:

Microdata Detail File (MDF; microdata with same schema as CEF)

Census TDA: Requirements and Properties II

Basic Structure of TDA



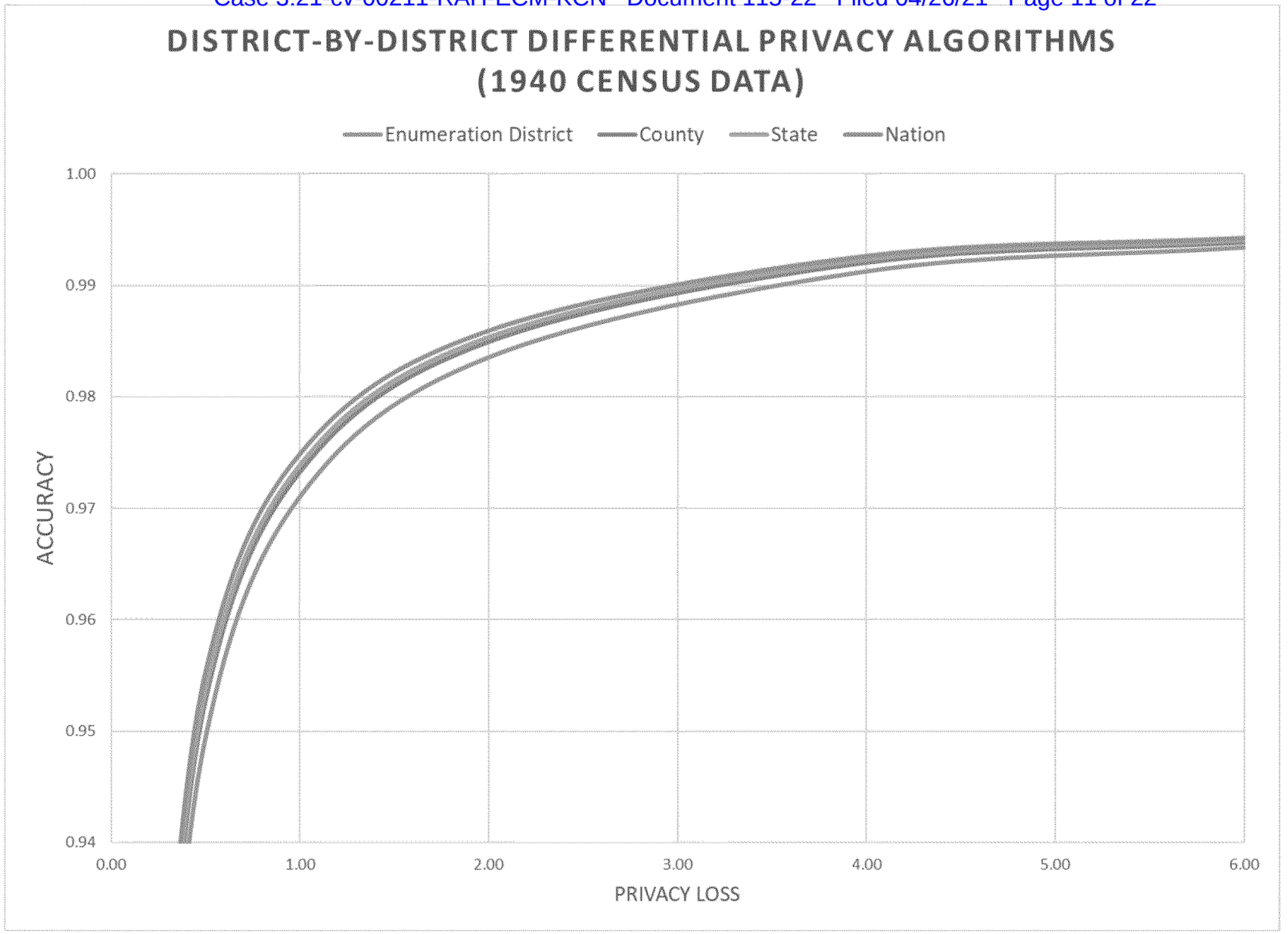
Microdata in TDA with Developing Invariants [1]

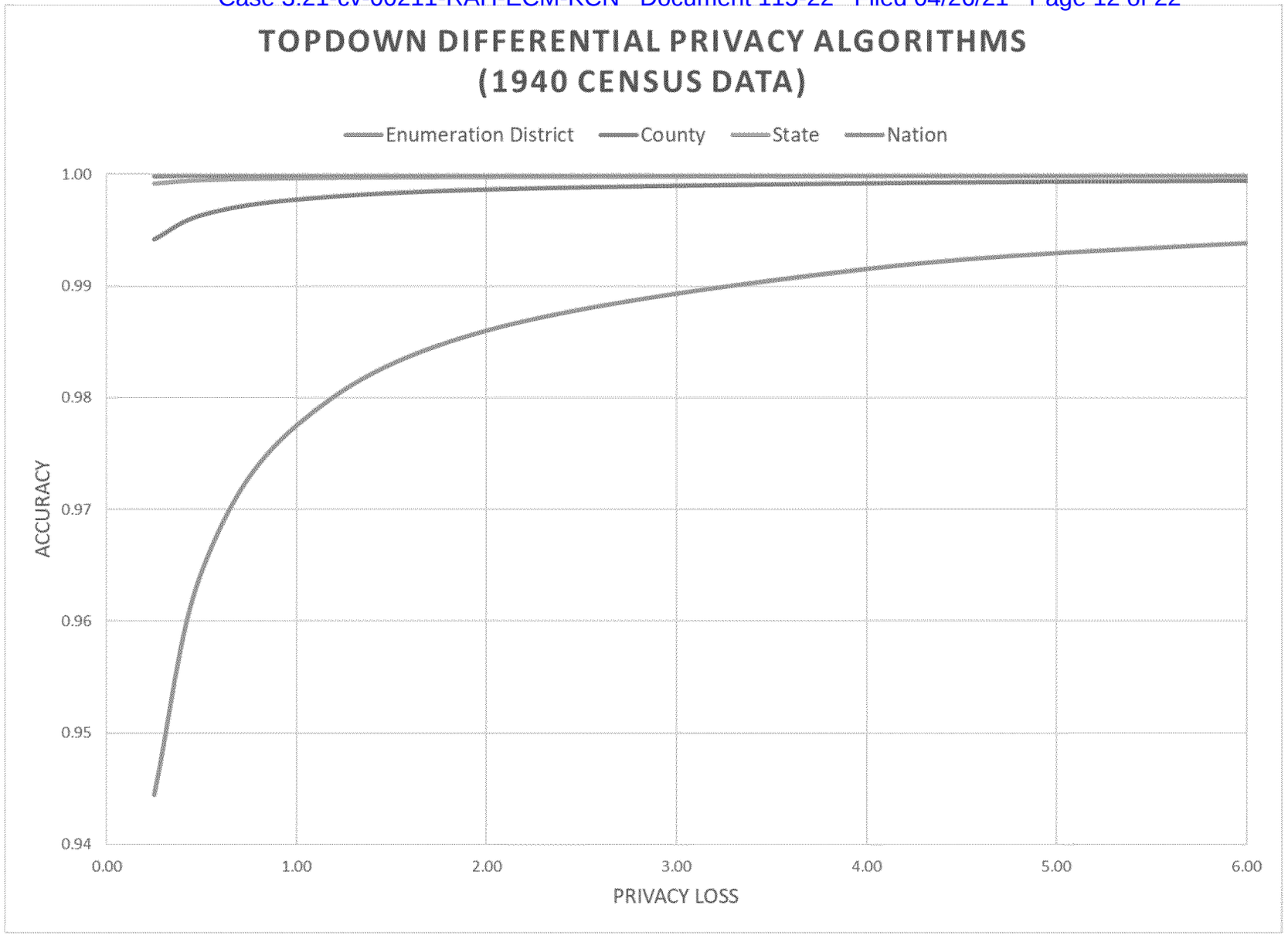
Microdata in TDA with Developing Invariants [2]

Microdata in TDA with Developing Invariants [3]

The Privacy-Accuracy Trade-off in TDA

- TDA is more complex than obvious competitors, but its error in a geounit does not increase with the number of contained Census blocks
- This is in particularly stark contrast to proceeding, e.g., Block-by-Block or District-by-District
- Most importantly, TDA yields increasing accuracy as the number of observations in a geographic unit increases





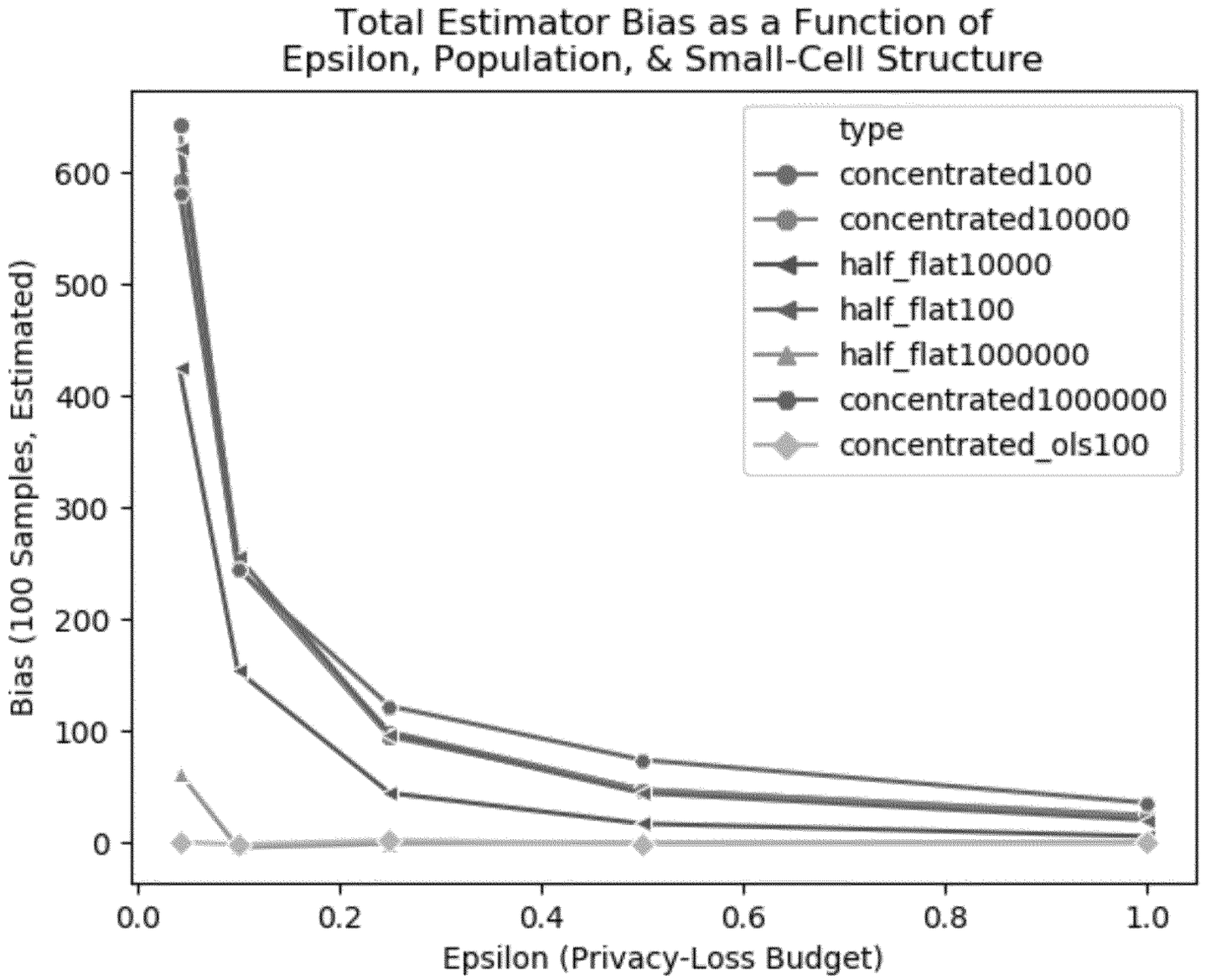
Beyond *1-TVD*: Outlier control in TDA

- In October 2019, the DAS team released a preliminary *Demonstration Data Product* illustrating the then-current development version of TDA: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html
- Global privacy-loss budget was 6 (4, for Persons; 2, for Households)
- Stakeholders analyzed the release, & shared a number of concerns, including:
  - Large outlier errors
  - Positive/upward bias
  - Large changes in Housing Vacancy / Occupancy rates
  - Systematic increase in bias & relative measures of error as tabulations get smaller

Since October 2019, the DAS team has worked to address these issues

- The standard suite of metrics calculated on the DAS was expanded, and increased emphasis placed on outlier behavior
- Worse relative error in tabulations with small expected counts is intrinsic; the privacy-accuracy trade-off is fundamental for these populations. Expanded metrics can help Census policy-makers reason about these trade-offs carefully
- A known fix was implemented for Vacancy/Occupancy rates: DP measurements were previously asymmetric (taken on Occupied but not Vacant Housing Units)
- The DAS team performed, and continues to perform, additional theoretical & computational work to improve our understanding of & implement strategies to ameliorate large outlier errors & upward bias

Positive Bias: A Numerical Example (1)



Positive Bias: Numerical Example (2)

Bias is driven mostly by "outlier" perturbations:

Perturbation Magnitude	Contribution to Total Bias
…	…
100	1.70E-11
101	1.69E-11
102	1.56E-11
103	1.65E-11
104	1.60E-11
105	1.63E-11
…	…
1050	10.93128571
1054	3.57
1057	3.504285714
1060	4.29
1069	4.095
1074	3.462
…	…

Overall Bias in Total estimator: 644.218

Theorem: Max Expected Error

The Max-Error Theorem leaves open:

- [1] Improvement of post-processing to reach the theorem's lower bound
- [2] Use of prior information to improve post-processing
- [3] The possibility of better lower bounds with adaptive measurements

The DAS team is targeting these openings

- [1] Use of statistical tests in post-processing to partition estimated counts into large & small subsets, and processing the large counts first to achieve OLS-like performance on this subset (Pursuing the lower bound)

- [2.A] Implementation of sequential NNLS optimization, targeting subsets of queries in multiple passes, with queries known to be less sparse (like total population) processed in earlier passes (Use of prior information)

- [2.B] Use of public historical data (from prior Census releases) to improve efficient estimation of sampling zeros in TDA histograms (Use of prior information)

- [3] Taking of DP measurements to take measurements in sequence, aggregating before measurements levels in later queries that are bounded near-0 by earlier queries (Adaptive measurements)

Current Status and Path Forward

- Preliminary implementations of improvements for each of [1]-[3] are done; empirical testing and refinement of these implementations is on-going

- The DAS team continues to examine other properties of TDA as well, and is currently studying its empirical large-epsilon behavior, with a focus on the properties of TDA's Rounding optimizations

- Final improvements & reports on those improved are expected in the next several months. Subsequently, hardening of the implementation will begin, in preparation for the final production runs of TDA

In case you have follow-up questions/comments…

Philip Leclerc

Mathematical Statistician

Center for Enterprise Dissemination-Disclosure Avoidance

Philip.Leclerc@census.gov

EXHIBIT 23

Cc: Tamara S Adams (CENSUS/ADDC FED)[Tamara.S.Adams@census.gov]; Barbara M LoPresti (CENSUS/DITD FED)[Barbara.M.LoPresti@census.gov]
To: Cathy A Ayoob (CENSUS/ADSD FED)[Cathy.A.Ayoob@census.gov]
From: Michael T Thieme (CENSUS/ADDC FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=768432D084ED41A6B1CF5785F7E56348-THIEME, MIC]
Sent: Fri 6/26/2020 3:10:43 AM (UTC)
Subject: Re: File integrity checks

Thanks, Cathy - I agree too.

Michael T. Thieme
Assistant Director for Decennial Census Programs, Systems and Contracts
U.S. Census Bureau
(301) 763-9062 (Office)
(301) 704-1594 (Mobile)
michael.t.thieme@census.gov

> On Jun 25, 2020, at 10:25 PM, Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov> wrote:
>
> Totally agree. Not sure why he has to create so much drama instead of just collaboratively workIng with all the team to properly test.
>
> Thanks,
> Cathy A Ayoob
> ADC for Quality and Enterprise Development Services
> Application Development & Services Division
> U.S. Census Bureau
>
> Office 301.763.4961 Room 3H149
> Cell 202.740.2858
> Fax 301.763.0333
>
> Cathy.a.ayoob@census.gov
>
> > On Jun 25, 2020, at 9:28 PM, Tamara S Adams (CENSUS/ADDC FED) <Tamara.S.Adams@census.gov> wrote:
> >
> > While in a classic sense, he has a point, but the technology used in soa reconstructs the chunks to which he refers.
> > That being said, we have no empirical evidence that there are issues. We've transferred a myriad of large files within and external to the bureau without issue.
> >
> > As for the iCade accusations, we had next to no outstanding issues last I checked.
> >
> > If we had tried to run our peak day optimization in ITE it would have failed. Ite isn't sized for that.
> >
> > He needs to have a file moved one time. I'm not sure why this much drama.
> >
> > **Tammy Adams**
> > Senior Advisor for Systems Operations and Optimization
> > MOJO Lead

Decennial Census Programs Directorate
U.S. Census Bureau

Office: 301-763-9258
Mobile: 202-580-5720
tamara.s.adams@census.gov

census.gov
Connect with us on Social Media

On Jun 25, 2020, at 9:00 PM, Michael T Thieme (CENSUS/ADDC FED) <Michael.T.Thieme@census.gov> wrote:

Can anyone tell me if Simson's claim about MFT is even really material to the success of DAS? With so much behind on that project and with all the help we are providing in good faith, I can't help but wonder if his MFT concerns, repeatedly made and answered, are only serving to distract from much more serious DAS problems. Thanks,
-Michael

Michael T. Thieme
Assistant Director for Decennial Census Programs, Systems and Contracts
U.S. Census Bureau
(301) 763-9062 (Office)
(301) 704-1594 (Mobile)
michael.t.thieme@census.gov

Begin forwarded message:

From: "John Maron Abowd (CENSUS/ADRM FED)" <john.maron.abowd@census.gov>
Date: June 25, 2020 at 7:32:44 PM EDT
To: "Cynthia Davis Hollingsworth (CENSUS/DCMD FED)" <cynthia.davis.hollingsworth@census.gov>, "Michael T Thieme (CENSUS/ADDC FED)" <Michael.T.Thieme@census.gov>, "Tamara S Adams (CENSUS/ADDC FED)" <Tamara.S.Adams@census.gov>, "Barbara M LoPresti (CENSUS/DITD FED)" <Barbara.M.LoPresti@census.gov>, "Quyen L Nguyen (CENSUS/CTO FED)" <quyen.nguyen@census.gov>, "Gerard Boudriault (CENSUS/DITD FED)" <Gerard.Boudriault@census.gov>
Cc: "Simson L Garfinkel (CENSUS/ADRM FED)" <simson.l.garfinkel@census.gov>, "Teresa Sabol (CENSUS/CED FED)" <teresa.sabol@census.gov>
Subject: Re: File integrity checks

Thanks Cynthia. And for the record, Paul Friday also identified this flaw more than a year ago and documented it in a memo that I communicated up the chain. I'm glad it's finally being addressed.

Best,

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

From: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Sent: Thursday, June 25, 2020 12:33 PM
To: Michael T Thieme (CENSUS/ADDC FED) <Michael.T.Thieme@census.gov>; Tamara S Adams (CENSUS/ADDC FED) <Tamara.S.Adams@census.gov>; Barbara M LoPresti (CENSUS/DITD FED) <Barbara.M.LoPresti@census.gov>; Quyen L Nguyen (CENSUS/CTO FED) <quyen.nguyen@census.gov>; Gerard Boudriault (CENSUS/DITD FED) <Gerard.Boudriault@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>
Subject: File integrity checks

Hi All,

We're continuing to encounter problems with SOA file transfers within ITE and have now begun testing within STAGE. I want to keep you in the loop since we've had previous discussions about file integrity.

In a separate email exchange between Simson and me, he documents what he believes to be the source of the SOA problem (which I'm copying below):

I'm working to do data transfer tests in both ITE and Staging.
It's my opinion that the data transfer errors are resulting from underlying flaws in the transfer protocol. Specifically:

- **Data is transferred in chunks.**
- **Each chunk has an offset and a length, but no validation code.**
- **If a chunk is lost, when the next chunk is received, the underlying implementation will execute a seek() in the destination file, which will result in a block of NULL bytes being inserted in the file.**

This is exactly the behavior we are seeing. Even if Stage has more capacity than ITE, the same flaw will be there, because it is a flaw in the *design* of the system, not in the scale. More capacity will simply make failures less likely. But a correctly design protocol will not result in data loss *ever*.

"Correct" is a technical term in computer science. The current MFT protocol is not correct. I noted this two years ago, when I first read the specification, and brought it to your attention at the time. We are now seeing the reliability impact of using a protocol that is not correct. Moving to Stage may make it more reliable, but it won't make the system correct. We will still have a chance of data

corruption when files are transferred.

Thanks,

Cynthia Davis Hollingsworth
Program Manager, 2020 Census Data Products and Dissemination
Decennial Census Management Division
U.S. Census Bureau
Office: 301.763.3655
iPhone: 202.253.6334
E-mail: cynthia.davis.hollingsworth@census.gov

From: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Sent: Thursday, June 25, 2020 11:15 AM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Vinod Raj Chavan Prakash (CENSUS/DCEO CTR) <vinod.raj.chavan.prakash@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>; Venkatsubramaniam Chandrasekharan (CENSUS/ADSD CTR) <venkatsubramaniam.chandrasekharan@census.gov>
Subject: Re: Updated dhcp mdf

Hi,
I unzipped random00.zip and got the bin file. I also did the following:

```
unzip -tq \*.zip
No errors detected in compressed data of random02.zip.
No errors detected in compressed data of random04.zip.
No errors detected in compressed data of random09.zip.
No errors detected in compressed data of random03.zip.
No errors detected in compressed data of random00.zip.
No errors detected in compressed data of random08.zip.
No errors detected in compressed data of random06.zip.
No errors detected in compressed data of random05.zip.
No errors detected in compressed data of random07.zip.
No errors detected in compressed data of random01.zip.
10 archives were successfully processed.
```

The file size was 9.8 GB each, and the transfer time for each file was about 4.75 to 5 hours.

Liza

Liza Hill, Chief, Decennial Tabulation Staff,
Decennial Information Technology Division, U.S. Census Bureau
Office 301.763.3582 liza.l.hill@census.gov
census.gov Connect with us on Social Media

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Thursday, June 25, 2020 10:55 AM
To: Vinod Raj Chavan Prakash (CENSUS/DCEO CTR) <vinod.raj.chavan.prakash@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>; Venkatsubramaniam Chandrasekharan (CENSUS/ADSD CTR) <venkatsubramaniam.chandrasekharan@census.gov>
Subject: Re: Updated dhcp mdf

Hi Vinod,

We need to stop using file length as an integrity check. We have seen on three occasions that files have been delivered with long runs of NULL characters. *Using file length as an integrity check will not catch this file corruption error that we are occasionally experiencing with the MFT.*

Do you have the ability to compute a cryptographic checksum on these files or, failing that, run the 'unzip -t' command on them? A cryptographic checksum will give us 160-bit integrity, which is the minimum we should be using. 'unzip -t' will give us a 32-bit integrity check. File length gives us zero bits of integrity.

Simson

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®

Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Vinod Raj Chavan Prakash (CENSUS/DCEO CTR) <vinod.raj.chavan.prakash@census.gov>
Sent: Thursday, June 25, 2020 10:40 AM
To: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>; Venkatsubramaniam Chandrasekharan (CENSUS/ADSD CTR) <venkatsubramaniam.chandrasekharan@census.gov>
Subject: Re: Updated dhcp mdf

Good Morning All,

I do see all the 10 files in CDL STG Inbound server. Please confirm the file size and also let me know if these needs to be archived for testing in STG .

<image.png>

Thanks,
Vinod

From: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>
Sent: Thursday, June 25, 2020 9:39 AM
To: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR)

<christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>; Venkatsubramaniam Chandrasekharan (CENSUS/ADSD CTR) <venkatsubramaniam.chandrasekharan@census.gov>; Vinod Raj Chavan Prakash (CENSUS/DCEO CTR) <vinod.raj.chavan.prakash@census.gov>
Subject: Re: Updated dhcp mdf

++ CDL team

Azizat (Abbey) Adalikwu (MSIT, CTFL, FAC-COR I) , IT Specialist
3H150A |ADSD | Enterprise Middleware Services Branch | HQ
US Census Bureau
O: 301-763-9713 | Cell -301-821-1658
Census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Thursday, June 25, 2020 9:29 AM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Good Morning Simson,

Unfortunately I cannot validate files on CDL servers.

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Thursday, June 25, 2020 9:20 AM
To: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Great. Liza should test file integrity with the unzip command. Bouna, can you test these files in CDL?

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Thursday, June 25, 2020 8:51 AM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED

CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Good Morning

All 10 files were delivered to CDL by 6:12 pm and TAB by 11:03 pm.

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Wednesday, June 24, 2020 5:59 PM
To: Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Okay, I just sent 10 zip files, each about 10GB in size.
I'm sending ZIP files so that you can verify the CRC32 with the unzip -t command.

Correlation ID DAS1593035849

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Wednesday, June 24, 2020 4:55 PM
To: Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Yea! Stand by for more...

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>
Sent: Wednesday, June 24, 2020 4:38 PM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>

Simson

The file has been delivered to both Tabulation and CDL (24-JUN-20 04.31.05.333000000 PM and 24-JUN-20 04.31.02.512000000 PM respectively).

Thanks,
Chris Horton

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Wednesday, June 24, 2020 4:31 PM
To: Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Sent with correlationID DAS1593030656

```
wrong-STAGING-MASTER:hadoop@/mnt/gits/das-vm-config $ python
soa/das_mft.py hello_world.txt --send2 --timestamp --debug
2020-06-23 17:39:28        20 hello_world.txt
Correlation ID:  DAS1593030656
Send file:  hello_world.txt
2020-06-24 16:30:57,330 connectionpool.py:203 (_new_conn) Starting
new HTTP connection (1): 169.254.169.254
2020-06-24 16:30:57,333 connectionpool.py:203 (_new_conn) Starting
new HTTP connection (1): 169.254.169.254
2020-06-24 16:30:57,350 connectionpool.py:735 (_new_conn) Starting
new HTTPS connection (1): kms.us-gov-west-1.amazonaws.com
headers:    {'content-type': 'text/xml', 'Authorization': 'Basic ='}
data:
 <soapenv:Envelope
xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/" >
  <soapenv:Header
xmlns:wsa="http://www.w3.org/2005/08/addressing">
  </soapenv:Header>
  <soapenv:Body>
```

```
      <v2:ProcessManagedFileTransferCBM
        versionID="string"

xmlns:v1="http://esoa.census.gov/soa/CanonicalModel/Core/Common
/V1"

xmlns:v2="http://esoa.census.gov/soa/CanonicalModel/Core/CDM/Ma
nagedFileTransferCDM/V2">
    <v1:CBMHeader>
     <v1:Sender description="string">DAS</v1:Sender>
     <v1:TargetList>
      <v1:Target description="string">CDL</v1:Target>
      <v1:Target description="string">TABULATION</v1:Target>
     </v1:TargetList>
    </v1:CBMHeader>
    <v2:ProcessManagedFileTransfer>
     <v2:CorrelationID>DAS1593030656</v2:CorrelationID>
     <v2:FileIdentifier>hello_world.txt</v2:FileIdentifier>
     <v2:TargetExtensionList>
      <v2:TargetExtension target="CDL">
       <ns1:extensionElement
          elementName="fileType"
          elementType="String"

xmlns:ns1="http://esoa.census.gov/soa/CanonicalModel/Core/Commo
n/V1">mdf</ns1:extensionElement>
       </v2:TargetExtension>
     </v2:TargetExtensionList>
    </v2:ProcessManagedFileTransfer>
  </v2:ProcessManagedFileTransferCBM>
 </soapenv:Body>
</soapenv:Envelope>

2020-06-24 16:30:58,246 connectionpool.py:735 (_new_conn) Starting
new HTTPS connection (1): kms.us-gov-west-1.amazonaws.com
Sent with Correlation ID:  DAS1593030656
SOA ERROR status:    200
SOA ERROR headers:   {'Date': 'Wed, 24 Jun 2020 20:30:59 GMT',
'Content-Type': 'text/xml; charset=utf-8', 'Content-Length': '0',
'Connection': 'keep-alive', 'X-ORACLE-DMS-ECID': '71c5ab20-a5cb-4757-
9d25-08c7b3eaf338-000790a2', 'X-ORACLE-DMS-RID': '0', 'messageID':
'afc4259.13d92d28.N26.172e1ae6ee0.c0a'}
------------------------------------------------------------
wrong-STAGING-MASTER:hadoop@/mnt/gits/das-vm-config $
```

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>
Sent: Wednesday, June 24, 2020 4:17 PM
To: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Confirmed-- The changes have been implemented in Stage.

Catherine Abad-Santos PMP, CSM, CSPO
Application Development & Services Division (ADSD)
U.S. Census Bureau
Catherine.S.Abad-Santos@census.gov

From: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>
Sent: Wednesday, June 24, 2020 4:05 PM
To: Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Cathy,
Did we finish the DAS deployment to the stage environment ?

Azizat (Abbey) Adalikwu (**MSIT, CTFL, FAC-COR I)** , IT Specialist

3H150A | ADSD | Enterprise Middleware Services Branch | HQ
US Census Bureau
O: 301-763-9713 | Cell -301-821-1658
Census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>
Sent: Wednesday, June 24, 2020 12:56 PM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Deployment for DAS routing change in STAGE (CR4306451 - CRQ45104) has been scheduled today at 2pm. It should be completed by 3pm.

Catherine Abad-Santos PMP, CSM, CSPO
Application Development & Services Division (ADSD)
U.S. Census Bureau
Catherine.S.Abad-Santos@census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Wednesday, June 24, 2020 12:13 PM
To: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED)

<teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Great. Let me know when the bucket name is corrected and I'll initiate another test! And then I'll send 500GB of data!

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Wednesday, June 24, 2020 11:22 AM
To: Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Catherine Susana Abad-Santos (CENSUS/ADSD CTR) <catherine.s.abad-santos@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

+ Cathy to give a status of the deployment of the config changes to stage.

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Wednesday, June 24, 2020 9:34 AM
To: Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Cynthia Davis Hollingsworth

(CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Damien,

We have the wrong bucket name configured in SOA. We need to make a change to our configuration to point to the right bucket name.

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>
Sent: Wednesday, June 24, 2020 9:24 AM
To: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Bouna,

Why would you change the bucket name? For one thing, it is named according to convention. Secondly, I see 13 places in the CloudFormation templates creating the environment in Staging that

reference the existing name. Why not just point to the bucket that actually exists?

Damian Anderson, Senior AWS Engineer, Contractor

DCEO/2020 Census Technical Integrator Program
U.S. Census Bureau
Mobile: 301-728-0919
Census.gov | 2020census.gov | @uscensusbureau

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Wednesday, June 24, 2020 9:20 AM
To: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Damien,

We will submit a ticket to change the bucket name. Since this is stage we have to go thru the CR process.

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Sent: Wednesday, June 24, 2020 9:08 AM
To: Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED

CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Thanks Joe.

@Simson L Garfinkel (CENSUS/ADRM FED) - you're good to go.

Thanks,

Cynthia Davis Hollingsworth
Program Manager, 2020 Census Data Products and Dissemination
Decennial Census Management Division
U.S. Census Bureau
Office: 301.763.3655
iPhone: 202.253.6334
E-mail: cynthia.davis.hollingsworth@census.gov

From: Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>
Sent: Tuesday, June 23, 2020 9:39 PM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Yes, Staging is ATO'ed for T13 Data.

Joseph Cortez, MS, MBA, Infrastructure Account Manager, Contractor

DCEO/2020 Census Technical Integrator Program

U.S. Census Bureau

M: 571.327.7616
Census.gov | 2020census.gov | @uscensusbureau

Planned PTO:

June 26 - Full Day
July 6 - Full Day
July 20-24 - One Week
August 7 - Full Day
August 14 - Full Day.

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Tuesday, June 23, 2020 6:19 PM
To: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

To anyone who feels qualified to answer.

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Sent: Tuesday, June 23, 2020 6:08 PM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED)

<Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Hi Simson,

To whom are you addressing your question?

@ Joe/TI - isn't Staging an environment approved for T13 data?

Thanks,

Cynthia Davis Hollingsworth
Program Manager, 2020 Census Data Products and Dissemination
Decennial Census Management Division
U.S. Census Bureau
Office: 301.763.3655
iPhone: 202.253.6334
E-mail: cynthia.davis.hollingsworth@census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Tuesday, June 23, 2020 5:53 PM
To: Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

I now need a CEF in the Staging environment. The easiest way for me to get it there is to transfer it to the management bucket, and then from the management bucket to the Staging bucket. The CEF is Title

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Tuesday, June 23, 2020 5:49 PM
To: Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

I just sent a single test file. When it is received, let me know, and I will send a big file.

The test file is called hello_world.txt. It is sent with correlation ID 'stage1'

```
wrong-ITE-MASTER:hadoop@/mnt/gits/das-vm-config $ python
soa/das_mft.py --send2 hello_world.txt --correlationID stage1 --debug
2020-06-23 17:39:28        20 hello_world.txt
Correlation ID:  stage1
Send file:  hello_world.txt
2020-06-23 17:47:50,333 connectionpool.py:203 (_new_conn) Starting
new HTTP connection (1): 169.254.169.254
2020-06-23 17:47:50,336 connectionpool.py:203 (_new_conn) Starting
new HTTP connection (1): 169.254.169.254
2020-06-23 17:47:50,353 connectionpool.py:735 (_new_conn) Starting
new HTTPS connection (1): kms.us-gov-west-1.amazonaws.com
headers:   {'content-type': 'text/xml', 'Authorization': 'Basic
U1RHX0RBUzpAQEBHZlI0NG00MzhqZmNnVjl0c3JGZWVlZ3gqKio='}
data:
 <soapenv:Envelope
xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/" >
  <soapenv:Header
xmlns:wsa="http://www.w3.org/2005/08/addressing">
  </soapenv:Header>
```

```
<soapenv:Body>
  <v2:ProcessManagedFileTransferCBM
    versionID="string"

xmlns:v1="http://esoa.census.gov/soa/CanonicalModel/Core/Common
/V1"

xmlns:v2="http://esoa.census.gov/soa/CanonicalModel/Core/CDM/Ma
nagedFileTransferCDM/V2">
   <v1:CBMHeader>
    <v1:Sender description="string">DAS</v1:Sender>
    <v1:TargetList>
     <v1:Target description="string">CDL</v1:Target>
     <v1:Target description="string">TABULATION</v1:Target>
    </v1:TargetList>
   </v1:CBMHeader>
   <v2:ProcessManagedFileTransfer>
    <v2:CorrelationID>stage1</v2:CorrelationID>
    <v2:FileIdentifier>hello_world.txt</v2:FileIdentifier>
    <v2:TargetExtensionList>
     <v2:TargetExtension target="CDL">
      <ns1:extensionElement
        elementName="fileType"
        elementType="String"

xmlns:ns1="http://esoa.census.gov/soa/CanonicalModel/Core/Commo
n/V1">mdf</ns1:extensionElement>
      </v2:TargetExtension>
    </v2:TargetExtensionList>
   </v2:ProcessManagedFileTransfer>
  </v2:ProcessManagedFileTransferCBM>
 </soapenv:Body>
</soapenv:Envelope>

2020-06-23 17:47:51,218 connectionpool.py:735 (_new_conn) Starting
new HTTPS connection (1): kms.us-gov-west-1.amazonaws.com
Sent with Correlation ID:  stage1
SOA ERROR status:    200
SOA ERROR headers:   {'Date': 'Tue, 23 Jun 2020 21:47:52 GMT',
'Content-Type': 'text/xml; charset=utf-8', 'Content-Length': '0',
'Connection': 'keep-alive', 'X-ORACLE-DMS-ECID': '3d557cd7-8cf8-427f-
ba16-148f3ff8e54d-000750bf', 'X-ORACLE-DMS-RID': '0', 'messageID':
'afc425d.N70d106dd.41.172e1ba38f8.Ne63'}
```

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>
Sent: Tuesday, June 23, 2020 8:42 AM
To: Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Thank you Damian

Joseph Cortez, MS, MBA, Infrastructure Account Manager, Contractor

DCEO/2020 Census Technical Integrator Program
U.S. Census Bureau

M: 571.327.7616
Census.gov | 2020census.gov | @uscensusbureau

Planned PTO:

June 26 - Full Day
July 6 - Full Day
July 20-24 - One Week
August 7 - Full Day
August 14 - Full Day.

From: Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>
Sent: Tuesday, June 23, 2020 8:42 AM
To: Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR)

<christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Joe: The incident is already assigned to me and I am working on it.

INC000000561871: Cannot launch cluster in staging - insufficient EC2 permission

Damian Anderson, Senior AWS Engineer, Contractor

DCEO/2020 Census Technical Integrator Program
U.S. Census Bureau
Mobile: 301-728-0919
Census.gov | 2020census.gov | @uscensusbureau

From: Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>
Sent: Tuesday, June 23, 2020 8:28 AM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Good Morning Simson,

I have alerted the Cloud Team of the situation.

Damian,
Can we discuss and escalate?

Joseph Cortez, MS, MBA, Infrastructure Account Manager, Contractor

DCEO/2020 Census Technical Integrator Program
U.S. Census Bureau

M: 571.327.7616
Census.gov | 2020census.gov | @uscensusbureau

Planned PTO:

June 19 - Half Day
June 26 - Full Day
July 6 - Full Day
July 20-24 - One Week
August 7 - Full Day
August 14 - Full Day.

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Tuesday, June 23, 2020 8:19 AM
To: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Joseph Cortez (CENSUS/DCEO CTR) <joseph.cortez@census.gov>; Damian Joseph Anderson (CENSUS/DCEO CTR) <damian.j.anderson@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Hi. We have not started testing in Stage.
+ @Joseph Cortez (CENSUS/DCEO CTR)
+ @Damian Joseph Anderson (CENSUS/DCEO CTR)

It appears that there is a configuration problem in stage, as a result of a recent effort to tighten up security. It is not currently possible to launch DAS clusters in Stage.

We do not have an ETA on having this resolved, but I expect that it will be resolved soon.

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®

Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>
Sent: Tuesday, June 23, 2020 7:39 AM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Good Morning Bouna,
Did we start running this test in Stage yet ? If we have not started yet when do we plan on starting ?

Azizat (Abbey) Adalikwu (**MSIT, CTFL, FAC-COR I)** , IT Specialist
3H150A |ADSD | Enterprise Middleware Services Branch | HQ
US Census Bureau
O: 301-763-9713 | Cell -301-821-1658
Census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Monday, June 22, 2020 5:27 PM
To: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob

(CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

We have not considered compressing these files. We are happy to do so. We could compress them as ZIP64 files so that you would get the checksum as well and send a single ZIP file called MDF20.zip

---

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Monday, June 22, 2020 11:51 AM
To: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Thanks Liza, please lets us know what you find. Also just out of curiosity have we considered compressing these files to shorten the transfer time?

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Sent: Monday, June 22, 2020 11:27 AM
To: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED)

<teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Hi Bouna,
I sent an email for this inquiry, and I don't know if I will hear back in time for your test. Based on previous experiences with data transfer, I believe your plan will work (50-10GB files).

Liza

Liza Hill, Chief, Decennial Tabulation Staff,
Decennial Information Technology Division, U.S. Census Bureau
Office 301.763.3582 liza.l.hill@census.gov
census.gov Connect with us on Social Media

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Monday, June 22, 2020 8:33 AM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Good Morning Simson,

Since this is first time we are testing in stage, i suggest we send 1 test file first and validate the file is readable by TAB before sending the rest of the files. Also we need to have discussion with TAB's Infra team as to how many concurrent sftp connections they have handle. I believe CDL can have 6 concurrent sftp connections per the testing we've previously done with them. We have to find out what that limit is for TAB. That should drive how many concurrent files we can transfer to TAB.

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Monday, June 22, 2020 8:26 AM
To: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Hi. I will not be ready to start the test until at last 3pm. I have training from 9am-noon and then meetings until 3.
I am able to start the ITE test then. I am hoping to be able to start the Staging test then.

I need to know two things:

1. Do you want me to send test files, or the actual MDF repeatedly?
2. Do you wish me to send all 500GB at the same time, or wait for each set to be received and acknowledged before sending the next?

I will be sending to both Tabulation and CDL, and I will send every file with a different name. In total, I am expecting to send 50 files of 10GB each.

My recommendation is to send test files, not the actual MDF (since it is Title 13), and to send all of the files at once, for maximum stress.

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE > 2020census.gov

From: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Sent: Friday, June 19, 2020 5:28 PM
To: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Hi Bouna,
We are ready and Monday morning will work. Thanks.

Liza

Liza Hill, Chief, Decennial Tabulation Staff,
Decennial Information Technology Division, U.S. Census Bureau
Office 301.763.3582 liza.l.hill@census.gov
census.gov Connect with us on Social Media

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Friday, June 19, 2020 5:14 PM
To: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Thank Simson and Liza,

If everybody is ready to run the test in stage I propose we do that Monday morning. We will monitor it on our side.

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Sent: Friday, June 19, 2020 4:21 PM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Hi,
We have close to 2TB of storage in that file system, and we can allocate 500 GB for this test. After the test, we will need to remove the files and free up storage.

Thanks,

Liza

Liza Hill, Chief, Decennial Tabulation Staff,
Decennial Information Technology Division, U.S. Census Bureau
Office 301.763.3582 liza.l.hill@census.gov
census.gov Connect with us on Social Media

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Friday, June 19, 2020 3:27 PM
To: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Cc: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED)

<cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Hi Bouna,

I have a theory of why this is happening.

My theory is that at some point in your infrastructure the large file is being transmitted as many segments. I think that when each segment is received, the receiving program does a fseek() and then a write(). I think that a segment is being lost in transmission. But when the next segment is received, I think that the seek() happens off the end of the file. UNIX semantics is would then NULL-fill the skipped region, which is precisely the behavior that you are seeing.

We can initiate as many file transfers as you wish. We can initiate them from either ITE or Stage or both. But Liza needs to have space to receive them! Let me know if you want us to send 100GB, 500GB, or 1TB, and we will send a series of 10GB files. If you wish, I can make them all different.

Simson

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality

U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

> On Jun 19, 2020, at 1:31 PM, Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov> wrote:
>
> Hi Simson,
>
> We looked into our traces and logs and didnt see any issues. We would like to have DAS and TAB's help further debug this issue. We would like DAS to initiate a few transfers in ite with unique file names, may be append the date and time to the file, and monitor the transfers. Please let us know when we can work on this.
>
> Best Regards
>
> Bouna Sall
> Application Development & Services Division (ADSD)

US Census Bureau
Office x31393
bouna.sall@census.gov

From: Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>
Sent: Friday, June 19, 2020 9:35 AM
To: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

Good Morning Simson,

We are looking into the traces and logs to figure our what happened. I will update this threads with our findings.

Best Regards

Bouna Sall
Application Development & Services Division (ADSD)
US Census Bureau
Office x31393
bouna.sall@census.gov

From: Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>
Sent: Friday, June 19, 2020 7:36 AM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR)

<christopher.r.horton@census.gov>; Jaime T Castello (CENSUS/ADSD FED) <Jaime.T.Castello@census.gov>; Cathy A Ayoob (CENSUS/ADSD FED) <Cathy.A.Ayoob@census.gov>; Luther Coleman McGinty (CENSUS/ADSD CTR) <luther.coleman.mcginty@census.gov>
Subject: Re: Updated dhcp mdf

+++Cathy, Jaime, and Luke

Azizat (Abbey) Adalikwu (MSIT, CTFL, FAC-COR I) , IT Specialist
3H150A |ADSD | Enterprise Middleware Services Branch | HQ
US Census Bureau
O: 301-763-9713 | Cell -301-821-1658
Census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Thursday, June 18, 2020 7:54 PM
To: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>
Subject: Re: Updated dhcp mdf

Liza,
We did not send a file with 190 MB of NULLs at the beginning. This file was corrupted by the managed file transfer protocol. This is the exact same corruption that we have seen in the past.
This is why it is vital that you confirm the cryptographic check some on the files that we sent you before processing them. The system that we are using to transfer files is not reliable.

This is a serious issue that has now happened on three occasions. It is vital that we realize that files are being corrupted. We can't do anything about it. Management refuses to take this seriously. I have been warning about

problems with the MFT system for two years. But there are no changes that will be made now.

Therefore, you must check to check the checksums before processing.

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality

U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

> On Jun 18, 2020, at 6:48 PM, Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov> wrote:
>
> Hi Simson,
> Before I moved the files, I could not read the top of the files. Bouna from the SOA team tried to debug, and I put in a ticket for CSvD to help solve the problem. Below was what CSvD found:
>
> Liza,
>
> You had me look into why issues were occurring in a specific file, /data/tab10/soa-in/ite/2020TRR-DAS-061520/MDF10_PER.txt. Here are the results:
>
> I hexed dump the file, since there are issues running UNIX commands against the files.
>
> A **valid file** the first few bytes look like this:
>
> ```
> [root@tabgen9 2020TRR-DAS-061620]# xxd
> MDF10_PER.txt | head -10
> 0000000: 2320 636f 6d62 696e 6564 2046
> 7269 204a  # combined Fri J
> 0000010: 756e 2031 3220 3134 3a31 363a
> 3538 2032  un 12 14:16:58 2
> 0000020: 3032 300a 2320 696e 7075 7431
> 3a20 7333  020.# input1: s3
> 0000030: 3a2f 2f75 7363 622d 6465 6365
> 6e6e 6961  ://uscb-decennia
> 0000040: 6c2d 6974 652d 6461 732f 7573
> 6572 732f  l-ite-das/users/
> ```

```
0000050: 6865 696e 6530 3038 2f6f 6666
6963 6961  heine008/officia
0000060: 6c2d 7472 722f 636f 7272 6563
7465 642d  l-trr/corrected-
0000070: 6468 6370 2d75 732f 4d44 4631
305f 5045  dhcp-us/MDF10_PE
0000080: 525f 5553 2e74 7874 2020 7072
6566 6978  R_US.txt  prefix
0000090: 3a20 5553 0a23 2069 6e70 7574
323a 2073  : US.# input2: s
```

The file you are having you are having issues with, is padded with zeros and looks like this:

```
[root@tabgen9 2020TRR-DAS-061520]# xxd
MDF10_PER.txt | head -10
0000000: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000010: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000020: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000030: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000040: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000050: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000060: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000070: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000080: 0000 0000 0000 0000 0000 0000
0000 0000  ................
0000090: 0000 0000 0000 0000 0000 0000
0000 0000  ................
```

I did another command to see when the zero padding ended. The first printable bytes appear about 19MB+ into the file.

I bolded the byte count, since the fields run together at this point.

```
[root@tabgen9 2020TRR-DAS-061520]# xxd
MDF10_PER.txt | awk
'{if($10!="................"){print
$0}}' | awk
'{if($9!="................"){print
$0}}' | more
474a00000 7c37 3739 3134 357c 337c 3030
307c 3231  |779145|3|000|21
474a000107c32 7c32 387c 327c 3032 7c32
0a4d 5044  |2|28|2|02|2.MPD
474a000207c31 2e31 2e30 7c37 327c 3132
397c 3232  |1.1.0|72|129|22
474a000303035 3032 7c32 7c32 3030 307c
```

```
3737 3832  0502|2|2000|7782
474a000403935 7c33 7c30 3030 7c32 357c
327c 3238  95|3|000|25|2|28
474a000507c32 7c30 327c 320a 4d50 447c
312e 312e  |2|02|2.MPD|1.1.
474a00060307c 3732 7c31 3239 7c32 3230
3530 327c  0|72|129|220502|
474a00070327c 3230 3034 7c37 3738 3935
337c 337c  2|2004|778953|3|
474a000803030 307c 3230 7c32 7c33 317c
327c 3032  000|20|2|31|2|02
474a000907c32 0a4d 5044 7c31 2e31 2e30
7c37 327c  |2.MPD|1.1.0|72|
474a000a03032 357c 3230 3231 3030 7c33
7c33 3030  025|202100|3|300
474a000b0307c 3739 3534 3837 7c33 7c30
3030 7c32  0|795487|3|000|2
474a000c0357c 327c 3134 7c32 7c30 377c
320a 4d50  5|2|14|2|07|2.MP
```

In conclusion, the files you are having issues with have zeros padding the beginning of the files. Based on the valid file, you are processing the file as a text file. All OS's use some kind of End-Of-Line markers to delineate the lines in the file. Since this file is padded with zeros, the first line is over 19MB long. You will have a hard time running any commands against it.

Most likely this file that was created needs to be rebuilt, without zero padding in the beginning.

Liza Hill, Chief, Decennial Tabulation Staff,
Decennial Information Technology Division, U.S. Census Bureau
Office 301.763.3582 liza.l.hill@census.gov
census.gov Connect with us on Social Media

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Thursday, June 18, 2020 8:53 AM
To: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>
Cc: Cynthia Davis Hollingsworth (CENSUS/DCMD

FED) <cynthia.davis.hollingsworth@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>
Subject: Re: Updated dhcp mdf

Liza,
I am very concerned by your email message that the head command would not work and you could not read the files into SAS. Do you have the files that were transferred that were unusable? Also, did you verify the cryptographic hashes on the files?

From your email, it sounds like the files were corrupted in transit. It sounds like a large block of the file contents were replaced with NULL characters. we have seen this kind of file corruption in the past. It is very important that we document this file corruption for ADSD and DCMD.

+ @Teresa Sabol (CENSUS/CED FED)
+ @John A Fattaleh (CENSUS/CED FED)

I am beginning to think that we should move file corruption onto the risk register. It seems that we have a hard time getting files to reliably transfer to your environment.

Lisa, I think that it is vital that you use the openssl 'sha1' command to verify the integrity of these files before you attempt to import them into your programming environment. That is why we are sending you the HASHES file.

can you get me access to the files that were sent that you could not read?

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE >
2020census.gov

From: Christopher John Rivers (CENSUS/DCMD

CTR) <christopher.j.rivers@census.gov>
Sent: Tuesday, June 16, 2020 9:36 AM
To: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Cc: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Azizat Adalikwu (CENSUS/ADSD FED) <Azizat.Adalikwu@census.gov>; Bouna Sall (CENSUS/ADSD CTR) <bouna.sall@census.gov>; Christopher Robert Stephen Horton (CENSUS/ADSD CTR) <christopher.r.horton@census.gov>
Subject: Re: Updated dhcp mdf

Good morning Simson and Micah, the DAS to TAB SOA transfer in ITE failed late in the transfer. Only 9.4GB received for the Unit file. Liza has removed the files. Can DAS please initiate a new SOA transfer and provide a correlation ID. I'm including SOA team to track. Thanks!

Chris Rivers, IT Project Manager
Decennial Census Management Division/HQ
U.S. Census Bureau
O: (301) 763-3528 | M: (757) 581-8884
census.gov | @uscensusbureau
Shape your future. START HERE >
2020census.gov

From: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Sent: Tuesday, June 16, 2020 9:24 AM
To: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>
Cc: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Updated dhcp mdf

Hi Chris,
Using the head command just hang, and SAS couldn't read in the files.

Liza

Liza Hill, Chief, Decennial Tabulation Staff,
Decennial Information Technology Division, U.S. Census Bureau

Office 301.763.3582 liza.l.hill@census.gov
census.gov Connect with us on Social Media

From: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>
Sent: Tuesday, June 16, 2020 8:16 AM
To: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Cc: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Updated dhcp mdf

Good morning Liza, I reached out to Micah and he stated he was able to run the Head command. Can you please check again before I ask them to re-send. Thanks!

[6/16/2020 8:13 AM] Micah Whitney Heineck (CENSUS/CED CTR):
I am able to run head on the MDF10_PER.txt, MDF10_UNIT.txt, and MDF10_HASHES.txt files. No problem on my side

Chris Rivers, IT Project Manager
Decennial Census Management Division/HQ
U.S. Census Bureau
O: (301) 763-3528 | M: (757) 581-8884
census.gov | @uscensusbureau
Shape your future. START HERE >
2020census.gov

From: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>
Sent: Monday, June 15, 2020 9:08 PM
To: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>
Cc: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Fw: Updated dhcp mdf

Hi Chris,
Do you know who can transferred the files again? There is a problem with the headers where I can't use the head command. Using the tail command is fine. We've had this problem not long ago. Thanks.

Liza

Liza Hill, Chief, Decennial Tabulation Staff,
Decennial Information Technology Division, U.S.
Census Bureau
Office 301.763.3582 liza.l.hill@census.gov
census.gov Connect with us on Social Media

From: Cynthia Davis Hollingsworth
(CENSUS/DCMD FED)
<cynthia.davis.hollingsworth@census.gov>
Sent: Monday, June 15, 2020 8:29 PM
To: Liza L Hill (CENSUS/DITD FED)
<Liza.L.Hill@census.gov>; Simson L Garfinkel
(CENSUS/ADRM FED)
<simson.l.garfinkel@census.gov>; Philip Leclerc
(CENSUS/CDAR FED) <philip.leclerc@census.gov>;
Teresa Sabol (CENSUS/CED FED)
<teresa.sabol@census.gov>; Christopher John
Rivers (CENSUS/DCMD CTR)
<christopher.j.rivers@census.gov>; Jane H Ingold
(CENSUS/DCMD FED)
<Jane.H.Ingold@census.gov>; Jason Devine
(CENSUS/POP FED) <Jason.E.Devine@census.gov>
Cc: Micah Whitney Heineck (CENSUS/CED CTR)
<micah.w.heineck@census.gov>; Charles Mcavoy
Lease (CENSUS/DCEO CTR)
<charles.m.lease@census.gov>; John A Fattaleh
(CENSUS/CED FED) <john.a.fattaleh@census.gov>;
Pavel Zhuravlev (CENSUS/CES CTR)
<pavel.zhuravlev@census.gov>
Subject: Re: Updated dhcp mdf

+ Jason

Hi Liza,

Glad you received the files!

Yes, you should use this file to tabulate the data products.

Thanks,

Cynthia Davis Hollingsworth
Program Manager, 2020 Census Data Products and Dissemination
Decennial Census Management Division
U.S. Census Bureau
Office: 301.763.3655
iPhone: 202.253.6334
E-mail: cynthia.davis.hollingsworth@census.gov

From: Liza L Hill (CENSUS/DITD FED)
<Liza.L.Hill@census.gov>

Sent: Monday, June 15, 2020 6:02 PM
To: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>
Cc: Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Charles Mcavoy Lease (CENSUS/DCEO CTR) <charles.m.lease@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Pavel Zhuravlev (CENSUS/CES CTR) <pavel.zhuravlev@census.gov>
Subject: Re: Updated dhcp mdf

Hi Simson and Phil,
We received the files at 5:36 p.m. Thank you.

May I use the files to tabulate the P.L., Demographic Profiles, and DHC data?

Liza

Liza Hill, Chief, Decennial Tabulation Staff,
Decennial Information Technology Division, U.S. Census Bureau
Office 301.763.3582 liza.l.hill@census.gov
census.gov Connect with us on Social Media

From: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Sent: Monday, June 15, 2020 2:45 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>
Cc: Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Charles Mcavoy Lease (CENSUS/DCEO CTR) <charles.m.lease@census.gov>; John A Fattaleh

(CENSUS/CED FED) <john.a.fattaleh@census.gov>; Pavel Zhuravlev (CENSUS/CES CTR) <pavel.zhuravlev@census.gov>
Subject: Re: Updated dhcp mdf

I agree that we should update the TAR to indicate that the run did not fail.

Simson L. Garfinkel, Ph.D, CISSP®, CIPP®
Senior Computer Scientist for Data Access and Confidentiality
U.S. Census Bureau
O: 301.763.5361 | M: 202.836.2859
census.gov | @uscensusbreau
Shape your future. START HERE >
2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, June 15, 2020 10:05 AM
To: Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>
Cc: Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Charles Mcavoy Lease (CENSUS/DCEO CTR) <charles.m.lease@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Pavel Zhuravlev (CENSUS/CES CTR) <pavel.zhuravlev@census.gov>
Subject: Re: Updated dhcp mdf

Sure, I can give an update at the sprint planning meeting.

re: the TRR documents -- it seems reasonable to me to include updates on this latest, believed-successful DHCP-US run, but I don't think we should remove notes that the original run failed (not clear to me if this was the intent or not, from the conversation below). It is an important development fact that an error was accidentally introduced (via a typo, in a merge not reviewed by multiple people; i.e., using our original development approach) into the DAS engine code that caused invariants to

fail.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>
Sent: Monday, June 15, 2020 7:12 AM
To: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>
Cc: Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Charles Mcavoy Lease (CENSUS/DCEO CTR) <charles.m.lease@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>; Pavel Zhuravlev (CENSUS/CES CTR) <pavel.zhuravlev@census.gov>
Subject: Re: Updated dhcp mdf

Yes, please, Chris, it makes sense to update the documents you reference with the correct run of the MDF file. Thank you for your offer.

@Phil - this is awesome news! Do you mind doing a quick update at today's Sprint V planning meeting, so that everyone is aware that the invariant bug has been corrected.

@Micah and Pavel - thank you for correcting the bug and for running the correct file.

Teresa Sabol, FAC P/PM III
Center for Enterprise Dissemination
Research & Methodology Directorate

U.S. Census Bureau
O: 301-763-6845 | M: 202-308-3505
census.gov | @uscensusbureau

From: Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>
Sent: Friday, June 12, 2020 10:26 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>
Cc: Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>; Teresa Sabol (CENSUS/CED FED) <teresa.sabol@census.gov>; Charles Mcavoy Lease (CENSUS/DCEO CTR) <charles.m.lease@census.gov>; John A Fattaleh (CENSUS/CED FED) <john.a.fattaleh@census.gov>
Subject: Re: Updated dhcp mdf

Good evening Philip, thanks for the response. Yes, can DAS please deliver the DHCP MDF to Tabulation in ITE.

Hi John and Teresa, please let me know if you would like the TRR Entry Checklist, RTVM Test Case Status, TAR, Release Notes,
TEMP updated with the recent successful test cases for Monday (OD TRR 6/15/2020)? Including Charles Lease if he has any recommendations. Thanks!

Chris Rivers, IT Project Manager
Decennial Census Management Division/HQ
U.S. Census Bureau
O: (301) 763-3528 | M: (757) 581-8884
census.gov | @uscensusbureau
Shape your future. START HERE >
2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Friday, June 12, 2020 4:28 PM
To: Liza L Hill (CENSUS/DITD FED) <Liza.L.Hill@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Christopher John Rivers (CENSUS/DCMD CTR) <christopher.j.rivers@census.gov>; Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>

Cc: Micah Whitney Heineck (CENSUS/CED CTR) <micah.w.heineck@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Updated dhcp mdf

Hi all,

I believe Micah and Pavel on the das team have identified the cause of the errors in the dhcp mdf, have fixed them, and that Micah has produced a new, correct dhcp-us mdf. Is there interest from tab in us sending a corrected mdf containing this data for review?

Best,
Philip Leclerc

Sent from my iPhone

EXHIBIT 24

To: Ryan R Cumings (CENSUS/CED FED)[ryan.r.cumings@census.gov]; Michael B Hawes (CENSUS/CED FED)[michael.b.hawes@census.gov]
From: John Maron Abowd (CENSUS/ADRM FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=CB0EEE1CC6CA45CC948C0077899626C2-ABOWD, JOHN]
Sent: Mon 7/13/2020 8:16:20 PM (UTC)
Subject: Re: Hill questions on DAS

Thanks.

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

From: Ryan R Cumings (CENSUS/CED FED) <ryan.r.cumings@census.gov>
Sent: Monday, July 13, 2020 1:43 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>
Subject: Re: Hill questions on DAS

Hi John and Michael,

Sorry about misunderstanding you on this. I included the message below summarizing a features of spines that have been topics of discussion in our science meetings, why they have an impact on the final estimator, and ways to avoid these difficulties.

I am not sure who the target audience is for this, but, if they have experience with data analysis, I could also go into more detail on a variety of aspects. The examples I have in mind are the following.

1. All sparse model estimators have infinite minimax risk, which was proved by Leeb and Potscher (2008).
2. The original form of the Lasso estimator (Tibshirani, 1996) actually makes the link between this estimator and the parent to child consistency constraints that I hint at below more explicit. Specifically, Tibshirani's original formulation was a constraint on the sum of the absolute values of the elements of the estimator. Thus, the non-negative Lasso can be formulated with a summation constraint on the estimator itself; it is currently more common to formulate this estimator with the Lagrange multiplier of this constraint as a choice parameter directly rather than the constant of this constraint.
3. I could also discuss a few other possibilities to ameliorate problems associated with high fan-out, but I think the opinion of most members of the science team (including myself) is that these possibilities are unlikely to be feasible to implement given our time constraints.

Feel free to let me know if you'd like me to update the email below.

Thank you,
Ryan

Hello X,

The aspect of the geographic spine that has the largest impact on the accuracy of the microdata produced by DAS is fan-out. The fan-out of a geounit is its number of child geounits. For example, the fan-out of Iowa in our current spine is the total number of counties in Iowa, which is 99. One obvious impact of increasing one or more fan-out values of a spine is that this increases the total number of geounits, which implies that the number of counts produced by DAS is necessarily higher. In a way that is made precise by the database reconstruction theorem, privacy becomes more difficult to ensure

as either the number of counts being released, or their precision, is increased, so it is not entirely surprising that our previous tests have shown that accuracy is often decreasing in fan-out.

There are also several other, less fundamental, reasons why fan-out has an impact on the accuracy of the DAS count estimates. First, our method of ensuring that each detailed cell count of a parent can be decomposed as the sum of the corresponding detailed cell count for all of its children is equivalent to a specific choice of Lasso penalty parameter on these detailed cells of the children. The addition of Lasso penalties (Tibshirani, 1996) are one common method for sparse model estimation, so this may not sound problematic at first. However, the variance of this implicit penalty term is increasing in the number of children, which in turn increases the variance of counts for children of geounits with a high fan-out. Second, fan-out values that are too low can also be problematic in some cases. The most obvious example is when a geounit only has one child, in which case it is best to simply bypass the parent and reallocate all of its privacy-loss budget (PLB) to the child. Third, increasing fan-out will also most typically increases the sparsity of the true cell counts that are estimated by the DAS algorithm. Since we also impose non-negativity constraints on the count estimates, this has the effect of increasing the bias of the cell count estimates.

A combination of these impacts can lead to less precise count estimates after modifying the spine to ensure that it pass through certain entities, such as AIAN areas. One approach that we are planning on testing is to take as input both a spine that includes AIAN areas as well as initial PLB allocations for each geolevel and then update this spine and the PLB allocations in a preprocessing step of the DAS algorithm. This can be done using an approach called the high dimensional matrix mechanism, which provides estimates of the optimal PLB allocations for each geounit (McKenna et al., 2018). Note that this solution is actually more general than simply optimizing the structure of the spine itself because we can bypass geounits that are allocated a PLB of zero to recover the required hierarchical spine structure.

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Monday, July 13, 2020 9:13 AM
To: Ryan R Cumings (CENSUS/CED FED) <ryan.r.cumings@census.gov>; Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>
Subject: Re: Hill questions on DAS

Ryan,

Please send Michael Hawes a summary of the state of geo-spine options. I'm not sure what clarification you are referring to by Simson. We need to be able to explain the issues associated with moving AIAN tribal areas onto the spine. These include fan-out and sparsity. These are science issues more than engineering. Please do this today.

Thanks,

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

From: Ryan R Cumings (CENSUS/CED FED) <ryan.r.cumings@census.gov>
Sent: Monday, July 13, 2020 8:42 AM
To: Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>; John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Subject: Re: Hill questions on DAS

Hi John,

Is this in regards to clarifying the methods related to the alternative geographic spine that we are planning on implementing to the participants of the Group 2 CNSTAT meetings? I was under the impression that our final decision was

not to send this email (after Simson's clarification).

I'm sorry if I misunderstood our conversation. Would you like me to send you a draft of an email today that describes our plans for testing alternative geographic spines?

Thank you,
Ryan

From: Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>
Sent: Monday, July 13, 2020 8:03 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Ryan R Cumings (CENSUS/CED FED) <ryan.r.cumings@census.gov>
Subject: Re: Hill questions on DAS

Ryan,

Do you have the new information on the geographic spine that John referenced in his Wednesday email (below)?

Thanks,
-Michael

Michael B. Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology
U.S. Census Bureau
301.763.1960 (office)
202.669.9035 (mobile)
michael.b.hawes@census.gov

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Wednesday, July 8, 2020 1:49 PM
To: Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>
Subject: Re: Hill questions on DAS

More detail on geo spine is coming by Friday from Ryan. Ping him, if you don't see it. Thanks,

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

From: Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>
Sent: Wednesday, July 8, 2020 1:48 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Christopher J Stanley (CENSUS/OCIA FED) <christopher.j.stanley@census.gov>
Cc: Ali Mohammad Ahmad (CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>; Alan Lang (CENSUS/OCIA FED) <alan.lang@census.gov>
Subject: Re: Hill questions on DAS

I'm available then, as well.

Michael B. Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology
U.S. Census Bureau
301.763.1960 (office)
202.669.9035 (mobile)
michael.b.hawes@census.gov

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Wednesday, July 8, 2020 12:02 PM
To: Christopher J Stanley (CENSUS/OCIA FED) <christopher.j.stanley@census.gov>; Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>
Cc: Ali Mohammad Ahmad (CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>; Alan Lang (CENSUS/OCIA FED) <alan.lang@census.gov>
Subject: Re: Hill questions on DAS

I can make 7/17 (although I will have to miss a virtual lunch with my counterparts at the other stat agencies).

We never "decided" to modify the geographic spine. We did experiment with determining other feasible spines, including one where AIAN tribal areas were on the spine. It solved some problems but created others (also for the AIAN tabulations), so we tabled it until we could figure out how to deal with the other problems. Everything is really sparse below AIAN state-level geography, so putting them on the spine meant that all the lower-level geographic measurements were worse, not better). Agree that this is a challenging problem, and we do need a fix for it, but we don't have one yet.

Thanks,

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

From: Christopher J Stanley (CENSUS/OCIA FED) <christopher.j.stanley@census.gov>
Sent: Wednesday, July 8, 2020 11:56 AM
To: Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>; John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Cc: Ali Mohammad Ahmad (CENSUS/ADCOM FED) <ali.m.ahmad@census.gov>; Alan Lang (CENSUS/OCIA FED) <alan.lang@census.gov>
Subject: Hill questions on DAS

Coincidentally right after Michael's great briefing for the Secretary, I just received list of questions from a staff member for Senator Gary Peters (MI), the ranking member on the Senate Homeland Security and Governmental Affairs Committee. For the last couple months, Al and Tim have been doing a standing weekly briefing for key staff from the congressional committees with jurisdiction over the census. I got a bunch of questions on hiring, operations and DAS. She asked that we cover these this week, but I want to keep the focus on census hiring, comms, and operations as usual for this week and then cover DAS in a later session.

Michael and/or John, could you join Friday 7/17 at noon to discuss these questions? I could also explore setting a separate time for DAS if our standing time can't work for you. The meeting is a Skype/phone meeting. The Hill staff all use the

phone number and not Skype online, so there would not be a presentation, only a verbal briefing. I would appreciate your assistance.

Here are the questions, with a little bit of commentary by me in brackets:

DAS:

- Why did the Bureau decide to keep the PRR for data processing at the original date, rather than shift it later, given the 3-month delay before data processing, and to allow more time for perfecting the DAS? [*I assume she's mixing topics, and my guess is that production readiness review (PRR) is for the computer systems needed for your work and separate from perfecting DAS*]
- In the adjusted nationwide DAS, after the latest sprint (Microdata Files for released on 7/1), did error measures improve for AI/AN tribal areas and other small, rural, remote populations? We have heard from stakeholders that error rates were even worse for some areas than in the original 2019 DAS. How does the Bureau plan to correct this in time for data processing? [*She is clearly talking to NCAI and reviewing their recent letters to us.*]
- Why did the Bureau reverse its decision to treat tribal areas similarly to states and cities, by incorporate tribal geographies into the geographic spine of the DAS (giving them their own privacy budget allocation)? Did the Bureau test this new geographic spine before rejecting it?
- What alternatives is the Bureau currently considering to ensure accurate results in tribal geographies?
- Has the Bureau scheduled a tribal consultation session that focuses solely on the DAS, to ensure each tribe has sufficient input on this matter by itself? [*We have already had two, and I heard from Dee that there might be another DAS consultation later.*]
- Why is the Bureau relying on CNS to tabulate the microdata it releases, rather than creating a demonstration product? What current duties prevent the relevant Bureau staff from working on this?

Chris Stanley, Chief
Office of Congressional and Intergovernmental Affairs
U.S. Census Bureau
O: 301-763-4276 | M: 202-280-9678
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

EXHIBIT 25

To: Victoria Velkoff (CENSUS/ADDP FED)[Victoria.A.Velkoff@census.gov]
From: John Maron Abowd (CENSUS/ADRM FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=CB0EEE1CC6CA45CC948C0077899626C2-ABOWD, JOHN]
Sent: Tue 7/21/2020 1:27:24 PM (UTC)
Subject: Re: Large Epsilon Additional Runs for Review

We can discuss, if you want. Controlling the error introduced by the detailed query is devilishly hard. We discussed a set of tests designed to get populations and PL94-171 (max query dimension 256) asymptotically consistent and monotone with good results in epsilon 15-25 range. We can review those next, but they won't be ready until Thursday.

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

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Tuesday, July 21, 2020 8:08 AM
To: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Than I expected. But not than Phil. Talking to him now.

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

From: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>
Sent: Tuesday, July 21, 2020 8:05 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Subject: Fw: Large Epsilon Additional Runs for Review

So if I am reading this correctly, even with epsilon of 500, we are seeing errors larger than anticipated?

Victoria Velkoff, PhD
Associate Director for Demographic Programs
U.S. Census Bureau
o: 301-763-1372
Shape your future. START HERE >2020census.gov
census.gov | @uscensusbureau

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Tuesday, July 21, 2020 7:42 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

This set appears to use one of the multipass NNLS variants. Is that correct?

Yes, it uses the basic multipass NNLS (not Robert's OLS-improved variant).

Unless I misunderstand something, the algorithm is monotonic with this rounder, but even at PLB 500, we have substantial error at every geographic level. That can't be due to the DP measurements, which all have standard deviations < 1, so it has to be due to NNLS.

The budget split for this run was:

```
epsilon_budget_total=
%(epsilon)s
                    geolevel_budget_prop= 0.2, 0.2, 0.15, 0.15, 0.15, 0.15
                    detailedprop= 0.1
                    dpqueries= total, hhgq * votingage * numraces, hhgq * votingage * hispanic * numraces, hhgq * votingage *
                    hispanic * cenrace
                    queriesprop= 0.3, 0.15, 0.15, 0.3
                    L2_DPqueryPart0= total
                    L2_DPqueryPart1= hhgq * votingage * numraces
                    L2_DPquerypart2= hhgq * votingage * hispanic * numraces
                    L2_DPqueryPart3= hhgq * votingage * hispanic * cenrace
                    L2_DPqueryPart4= detailed
```

(https://github.ti.census.gov/CB-DAS/das_decennial/blob/b7625df532b6567c3166e7994c8dd3b3e8d21da1/configs/full_person/multiL2_singlePassRounder_RI_nested.ini#L73)

At epsilon of 500, this leaves the detailed query (as the worst-case example) with "local" budget of

```
>>> 500. * 0.1 * 0.15
7.5
```

That is expended on ~500K samples per geographic unit, which still contain substantial error, even just in the DP measurements:

```
>>> x, y = prng.geometric(p, 500000) - 1., prng.geometric(p, 500000) - 1.
>>> np.sum(np.abs(x - y))
23458.0
```

That is not true for some of the other statistics in use here. For example, hhgq * votingage * hispanic * cenrace has "local" epsilon:

```
>>> epsilon = 500. * 0.3 * 0.15
>>> epsilon
22.5
```

Which will typically have 0 error in a single geounit:

```
>>> x, y = prng.geometric(p, 8*2*2*63) - 1., prng.geometric(p, 8*2*2*63) - 1.
>>> np.sum(np.abs(x - y))
0.0
```

Though this isn't true at lower geographic levels. RI has ~25K Census blocks, for example, so at that local level error in DP measurements is still non-zero, despite the larger budget assigned to this query:

```
>>> x, y = prng.geometric(p, 8*2*2*63*25000) - 1., prng.geometric(p, 8*2*2*63*25000) - 1.
>>> np.sum(np.abs(x - y))
1267.0
```

That said, I suspect the detailed query may be inflating the error somewhat. Since it works over the floats, not the integers, multipass doesn't constrain later passes to exactly match query estimates from earlier passes; instead, it only requires they agree within a tolerance, which I have set to 5.0 by default:

https://github.ti.census.gov/CB-DAS/das_decennial/blob/29a2823973d90c4b724dc437c18620cff7fc4f83/programs/optimization/l2_dataIndep_npass_optimizer.py#L144
https://github.ti.census.gov/CB-DAS/das_decennial/blob/29a2823973d90c4b724dc437c18620cff7fc4f83/programs/optimization/l2_dataIndep_npass_optimizer.py#L176

As a result, the larger error in the detailed query can mildly deteriorate the other estimates in each geounit. And this effect can compound additively as we move down the geohierarchy -- introducing 5 extra error at the County level, then 5 more at the Tract_Group level, and so on (and this can happen within each scalar of each vectorized query, so it is worse for the 8*2*2*63 query than for the Total Pop query).

A few other things to note:

- multipass is likely to converge more slowly to 0 error than our original approach with a single simultaneous optimization (though Robert's OLS-improved multipass should help with this)
- there is room to modify the budget settings; this is still quite a preliminary run. We might try re-allocating some PLB away from total pop, and away from the top two geolevels
- I could try reducing the multipass tolerance a bit, although this will eventually induce instability
- currently, the NNLS solve and the Rounder target the same queries, and we recently learned that this new Rounder requires "hierarchically nested" queries, so I was forced to modify the NNLS queries to respect this property as well. This restriction can be lifted, though, which would allow us more freedom in the NNLS query specification (which may be to the benefit of convergence rate for our queries of interest)

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Monday, July 20, 2020 11:27 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Easiest way is to just turn on the viewer and rummage around.

Unless I misunderstand something, the algorithm is monotonic with this rounder, but even at PLB 500, we have substantial

error at every geographic level. That can't be due to the DP measurements, which all have standard deviations < 1, so it has to be due to NNLS. This set appears to use one of the multipass NNLS variants. Is that correct?

Crashed my Chrome, so I invoked my "first crash after 9pm = all done" rule. All done for today.

Thanks,

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

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, July 20, 2020 3:07 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

If you mean our (DAS team analysis module) usual error visualizations, John, I've been doing so; the script did not like my asking it to process a much larger number of parameter variations for usual, so I'm having to fill in combinations it missed somewhat piecemeal, but I've been documenting this in:

https://github.ti.census.gov/CB-DAS/das_decennial/issues/393#issuecomment-12534

The primary folder of interest is:

RM_SHARED:\dms-p0-992\CNSTAT_DDP_Improvement_Experiments_March_2020\asymptotic_epsilon_investigation\lecle301\nested_queries_multiL2_singlePassRounder\visualizations\singlePassRounder_VA

Images are best opened [A] as pngs and [B] in Chrome (or the point at which the volume is mounted moved down several sub-folders) (otherwise, Windows will complain about too-long path names)

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Monday, July 20, 2020 3:02 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Thanks. Could you please stage the salient ones to dms-p0-992?

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

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, July 20, 2020 2:23 PM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Hi Matt (+ Tori, John, Cynthia to cc),

We've made some notable progress on the large-epsilon runs, though we haven't tried any Nation-wide yet and we have some further tuning of budget parameters to do, and one more refinement of the method to implement.

If you (or others with access to/knowledge of the metrics scripts) have time, you may want to analyze the following VA and RI data (the bolded, underlined ones):

tacos-ITE-MASTER:hadoop@ip-10-252-47-18$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_multiL2_

PRE cnstatDdpSchema_multiL2_cellWiseRounder_nested_accuracyTest/
PRE cnstatDdpSchema_multiL2_cellWiseRounder_nested_accuracyTest_RI/
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest/
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest_-1./
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest_100/
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest_eps100_singleRounderPass/
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest_manyEps_singleRounderPass/
PRE cnstatDdpSchema_multiL2_singlePassRounder_nested_accuracyTest/
PRE cnstatDdpSchema_multiL2_singlePassRounder_nested_accuracyTest_RI/

Of the four, the two marked RI are, well, RI data, and the two not marked RI are (poorly named) VA data.

cellWiseRounder denotes our original rounder. singlePassRounder is one of the two new methods, which from what I can tell works as intended (largely -- maybe completely -- restoring monotonicity of accuracy in epsilon); it adds L1 penalty terms to the Rounder, in addition to the Rounder's usual detailed-cell terms.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, June 1, 2020 1:09 PM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>

Hey Matt,

Sort of -- we did *many* re-runs, and conducted a lengthy investigation (mostly documented here, if you're curious: https://github.ti.census.gov/CB-DAS/das_decennial/issues/264), but none of them would be useful for you all to analyze right now. Basically, we identified two issues that need to be fixed before it is probably worthwhile to re-run large *epsilon* runs for proper Analysis:

- a while back, to improve the security of our pseudo-random number generator, we switched from using standard *numpy* random distributions/samplers to using an Intel distribution of Anaconda that has an alternative library, *mkl_random*, for doing the same thing. Unfortunately, we didn't realize that *mkl_random* makes several non-obvious, unadvertised changes -- including, notably, that it behaves improperly (gives nonsensical errors) at degenerate scale parameters (including, specifically, that it yields arbitrary, large perturbations when fed a scale value of *1.0* for a Geometric distribution; this is the opposite of the way base *numpy* operates, which behaves as you'd expect it to in the continuous limit, i.e., yields negligible/zero noise at scale 1.0)
- as *epsilon* increases, the importance of the Rounder problem (which converts float-valued estimates to integer-valued estimates) increases. This is unfortunately expensive (in terms of accuracy), because the Rounder is not designed to be as statistically efficient as the main NNLS solves (a sacrifice made because the Rounder has to be structurally simpler to guarantee that it will be able to find an integer-valued solution at all)

The fix for the first problem is pretty simple, and we can implement it and guarantee that, eventually, large epsilon will give perfect accuracy.

But that's probably not enough; the second problem requires a little more work to fix, but is very important, because it can create "non-monotonicities": queries can get worse as epsilon increases (after post-processing, because the Rounder has become important), before eventually getting better (because even the Rounder is eventually perfect, if 0 noise is introduced). That can make the relationship of *epsilon* to accuracy more complicated than we want it to be, so we need to implement improvements to tackle this second problem before we re-generate "official" large-epsilon runs.

Well, unless we want epsilon so large that accuracy is near-perfect, anyway. That can be done with just the simple fix, but I think it is probably not sufficient.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Sent: Monday, June 1, 2020 12:33 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Hi Phil-

I hope you're well. Did you ever do a re-run of these large epsilon runs?

Thanks,
-Matt

Matthew Spence, Branch Chief
Foreign-Born Population Branch
Population Division
U.S. Census Bureau
o: 301-763-1033
census.gov | @uscensusbureau

Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, May 18, 2020 10:03 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Hi all,

I've completed initial Analyses on the large-epsilon runs, but the results look somewhat odd (implausibly high error). Would hold off on reviewing them until we can investigate.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Friday, May 15, 2020 10:33 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Large Epsilon Additional Runs for Review

Hi all,

After conferring with John and Dan K. on Wednesday, we elected to generate additional National runs at very large epsilons. Two new runs, using each of singlePassRegular (basic TopDown) and dataIndUserSpecifiedMultipass (basic multipass) and total *epsilon=100*, have completed and are available for analysis at:

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_National_dpQueries100_largeEpsRun/

and

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries100_largeEpsRun/

Note that, even at *epsilon=100*, non-zero noise was almost certainly introduced (because the budget is divided in 5 parts per level, for 7 levels, so the on-average local *epsilon* expenditure on a query is approximately *100/35* -- which is still large, but not so large as to expect 0 noise over billions of noisy samples).

Best,

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Saturday, May 2, 2020 10:05 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/EWD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP Wigglesum runs # 1-4 (Persons Universe) for CNSTAT Experts Meetings

The dpQueries5 wigglesum run has completed now as well. It is at:

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_TwoPassBigSmall_National_dpQueries5_AllChildFilter

(this is the same location as in the last email, but at the time the dpQueries5 s3 "directory" was only partially populated)

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Friday, May 1, 2020 3:47 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/EWD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>

Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Official CNSTAT DDP Wigglesum runs # 1-4 (Persons Universe) for CNSTAT Experts Meetings

Hi all,

Note changed thread title. The dpQueries1-4 Wigglesum runs are complete; their microdata is available in our s3 bucket at the underlined/bolded/italicized locations below:

brach-ITE-MASTER:hadoop@ip-10-252-47-189$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_TwoPassBigSmall_National_dpQueries

PRE ***cnstatDdpSchema_TwoPassBigSmall_National_dpQueries1_AllChildFilter/***
PRE cnstatDdpSchema_TwoPassBigSmall_National_dpQueries2_AllChildFilter/
PRE ***cnstatDdpSchema_TwoPassBigSmall_National_dpQueries2_AllChildFilter_attempt2/***
PRE cnstatDdpSchema_TwoPassBigSmall_National_dpQueries3_AllChildFilter/
PRE ***cnstatDdpSchema_TwoPassBigSmall_National_dpQueries3_AllChildFilter_attempt2/***
PRE ***cnstatDdpSchema_TwoPassBigSmall_National_dpQueries4_AllChildFilter/***
PRE cnstatDdpSchema_TwoPassBigSmall_National_dpQueries5_AllChildFilter/

The dpQueries5 Wigglesum run will complete sometime tomorrow morning (5/2/2020). For dpQueries2 and dpQueries3, please use the "attempt2" locations (as bolded above); the first two attempts for those budget settings failed because we ran out of gurobi licenses. (That is also the reason for the small delay in this email, and in delivery of dpQueries5 MDF data; as it turns out, we do not have enough gurobi licenses to execute 4 National runs simultaneously!)

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Tuesday, April 28, 2020 8:17 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

Standard DAS error visualizations have been prepared and downloaded to the dms-p0-992 folder. Currently copying them over to DAS_Collaboration. Four new folders were added in *Y:\CNSTAT_DDP_Improvements\preliminary_analyses\Persons*

cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries2_officialCNSTATrun
cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries3_officialCNSTATrun
cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries4_officialCNSTATrun
cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries5_officialCNSTATrun

As previously indicated, these differ from one another in that, although all have the same overall privacy-loss budget of 4.0, dpQueriesX assigned increasing proportions of the privacy-loss budget to the Total query as X gets larger. This should improve error on Total Population, but worsen error on other tabulations.

These should be compared with the original CNSTAT run (noting that some of the labels were out-of-order in the original National CNSTAT run Analysis), in sub-folder:

cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries1_officialCNSTATrun

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, April 27, 2020 3:23 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

Update: looks like the visualizations won't be ready this evening; all 4 analysis scripts are still running, but are in the middle of the PL94 tabulations (perhaps I should have used larger clusters).

I'll check on them again later tonight, but it seems likely they may have to run overnight.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Monday, April 27, 2020 10:35 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

Got it. Thanks. Planning to review this evening if the visualizations are ready.

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

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, April 27, 2020 10:33 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

p.s. Oops, I left a typo here -- this should say dpQueries***1***:

dpQueries2, 0.05 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries1_officialCNSTATrun.ini#L63 (this run is not new; it was the original Persons universe run, which we examined Matt's error metrics on last Wednesday)

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, April 27, 2020 10:32 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

Hi all,

Note changed thread title.

After review of error metrics for the # 1 CNSTAT Experts meeting run for the Persons universe last Wednesday, I was asked to generate additional runs, with increasing proportions of the overall privacy-loss budget [PLB] assigned to the Total Population query. I believe John and Tori will select one of these runs to use as the 'official Persons-universe run (or, if none of these seem acceptable, may ask for additional run(s)). These runs are complete, and can be found at:

```
abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-
das/users/lecle301/cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQuerie
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries1_officialCNSTATrun/
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries2_officialCNSTATrun/
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries3_officialCNSTATrun/
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries4_officialCNSTATrun/
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries5_officialCNSTATrun/
```

The budget sections of the configuration files show the PLB allocation and measurements taken for each of these runs:

dpQueries2, 0.05 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries1_officialCNSTATrun.ini#L63 (this run is not new; it was the original Persons universe run, which we examined Matt's error metrics on last Wednesday)

dpQueries2, 0.2 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/328209709cac43fbaf2c1062d8271812269056a3/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries2_officialCNSTATrun.ini#L63

dpQueries3, 0.3 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries3_officialCNSTATrun.ini#L63

dpQueries4, 0.5 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries4_officialCNSTATrun.ini#L63

dpQueries5, 0.75 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries5_officialCNSTATrun.ini#L63

Note that, as the proportion of PLB assigned to Total increases, we expect error generally to increase on the remaining queries/tabulations.

I am currently working on the standard DAS Analyses visualizations; they're not ready yet, but I will copy them to DAS Collaboration and update here when they are prepared (likely by COB today).

Best,

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, April 22, 2020 9:49 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP

FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 1 for CNSTAT Experts Meetings

Hi all,

I was asked in our 9 AM meeting today when the official Nation-wide CNSTAT Experts runs, & unpickled H1-only measurements, were delivered.

Please see the below email from 4/15/2020; it constituted that delivery.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, April 15, 2020 2:13 PM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Official CNSTAT DDP (re-)runs # 1 for CNSTAT Experts Meetings

Hi all (also, + Simson to cc),

Note re-named thread title.

The Units (H1-only) and Persons National runs (using the `DHCP_HHGQ` schema, which is the same histogram we relied on for the CNSTAT DDP release) for the next CNSTAT Experts meeting have both completed, and individual algorithm variants/trials have been selected by Tori and John as those we'll use for the next experts update. In addition, I've fixed the header de-duplication issue in the H1-only measurements, and unpickled the H1-only measurements for the official Units run.

The official data locations in s3 are as follows.

Persons MDF:

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries1_officialCNSTATrun/data-run1.0-epsilon4.0-DHCP_MDF.txt

H1-only Units MDF:

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_nat_H1_Only_withMeasurements_v8/MDF_UNIT-run1.0-epsilon0.5-H1.csv

H1-only unpickled pipe-delimited measurements text files (Occ/Vac count estimates, integer but can be negative; consistent, because there were only the two values measured; one file per geolevel; each file with header)**:**

Within s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_nat_H1_Only_withMeasurements_v8/noisy_measurements-eps0.5-run1/

The relevant files are:

application_1586267880679_0010-Block.csv
application_1586267880679_0010-Block_Group.csv
application_1586267880679_0010-Tract.csv
application_1586267880679_0010-Tract_Group.csv
application_1586267880679_0010-County.csv
application_1586267880679_0010-State.csv
application_1586267880679_0010-National.csv

I think this concludes this first pass-off of data from the DAS team to POP/DEMO for the experts meetings. Please let me know if there are questions/concerns/access issues, etc.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Thursday, April 9, 2020 7:07 PM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Subject: Re: data from recent Virginia runs

Hi everyone,

Adding Jason, Tori, John, Christine B. to this thread. (Please add anyone else who should have access to this data.)

In addition to sharing experimental MDFs (post-processed DP measurements) from a variety of algorithms, the DAS team was asked/had promised to provide DP measurements (before post-processing) for one of the H1-only runs. This will be a good starting point for understanding them, because the H1-only measurements are very simple (just two numbers per geounit). For POP/DEMO folks with access to our s3 bucket, I've provided this data for each geographic level in the first

PLB 0.007 H1-only run here:

abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_nat_H1_Only_withMeasurements_v8/noisy_measurements-eps0.007-run1/application_158626788 | grep .*\.csv

```
                           PRE application_1586267880679_0010-Block.csv/
                           PRE application_1586267880679_0010-Block_Group.csv/
                           PRE application_1586267880679_0010-County.csv/
                           PRE application_1586267880679_0010-National.csv/
                           PRE application_1586267880679_0010-State.csv/
                           PRE application_1586267880679_0010-Tract.csv/
                           PRE application_1586267880679_0010-Tract_Group.csv/
2020-04-09 18:36:42   application_1586267880679_0010-Block.csv
2020-04-09 18:39:20   application_1586267880679_0010-Block_Group.csv
2020-04-09 18:50:13   application_1586267880679_0010-County.csv
2020-04-09 18:55:08   application_1586267880679_0010-National.csv
2020-04-09 18:51:23   application_1586267880679_0010-State.csv
2020-04-09 18:43:26   application_1586267880679_0010-Tract.csv
2020-04-09 18:48:38   application_1586267880679_0010-Tract_Group.csv
```

I think the measurement files' structure should be self-explanatory, but let me know if there are questions. Also please let me know if you think you've found any errors in them, too, of course!

Two important notes:

- an unfortunate quirk remains -- I accidentally let the concatenation program duplicate the csv's header many times. So, when parsing these, you'll have to remove rows that look like: *DPQueryName|geocode|Vacant|Occupied*. There will be more than just one such row! (I can fix this early next week, but this caveat applies for anyone who uses this data before then.)
- the PLB in this case is *very* small, but it is also what I crudely estimate we expended on estimating the number of Occupied Housing Units in the CNSTAT DDP release (in case you were wondering where 0.007 came from). We also have runs for PLBs 0.1, 0.5, 2.0, 1.0, and 4.0, and I could similarly unwrap DP measurements for some of those runs, if there's interest

Also, for H1-only, these files are relatively small, so moving them to other secure locations should be relatively easy, if desired. (That will not be the case when we share the DP measurements for full runs; for DHC(-P), for example, the full measurement sets tend to be measurements in dozens of terabytes.)

Best,

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Thursday, April 9, 2020 4:57 PM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>

Subject: Re: data from recent Virginia runs

Hi all,

I've corrected the incorrect visualizations and replaced the copy in DAS_Collaboration.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Thursday, April 9, 2020 11:41 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Oh, an unfortunate update on the visualizations in *preliminary_analyses*: in the most recent runs, they computed *average signed error*, not *mean absolute error*, because of a change I made to improve scalability (not noticing that I had lost a *.abs* in the process).

They'll have to be re-computed with the *.abs* fixed.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Sent: Thursday, April 9, 2020 8:28 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Thanks Matt and Philip! I'll submit the DIRT ticket this morning.

Luke

Luke T. Rogers, PhD
Chief, Population Estimates Branch
Population Division
U.S. Census Bureau

O: 301-763-7147
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Thursday, April 9, 2020 8:24 AM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

That's right, Matt, yep.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Sent: Thursday, April 9, 2020 8:04 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Hi Luke -

I did complete a DIRT ticket request for \\it171oafs-oa03.boc.ad.census.gov\DEMO_SHARE\DAS Collaboration yesterday afternoon. I believe \CNSTAT_DDP_Improvement_Experiments\preliminary_analyses\ is nested within DAS Collaboration. (Phil, is that right?)

Matthew Spence, Branch Chief
Foreign-Born Population Branch
Population Division
U.S. Census Bureau
o: 301-763-1033
census.gov | @uscensusbureau

Shape your future. START HERE > 2020census.gov

From: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Sent: Thursday, April 9, 2020 6:40 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>

Subject: Re: data from recent Virginia runs

Thank you Philip! Excited to see your team's results.

I would like access to \CNSTAT_DDP_Improvement_Experiments\preliminary_analyses\ .

Matt - did you just request access through DIRT?

Luke

Luke T. Rogers, PhD
Chief, Population Estimates Branch
Population Division
U.S. Census Bureau
O: 301-763-7147
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, April 8, 2020 5:30 PM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Hi all,

We have more data, now. For the Persons universe, the version7 and v7 runs here are good for analysis:

```
abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_ | grep [4,6]_v.*7
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries4_version7/
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries6_v7/
```

We also have Housing Unit (just Occupied/Vacant, per previous conversations) runs at various epsilons. They can be found here:

```
hazan-ITE-MASTER:hadoop@ip-10-252-46-63$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_na_H1_Only_withMeasurements_v8/ | grep .*\.csv

2020-04-07 16:18:23  138923937 MDF_UNIT-run1.0-epsilon0.007-H1.csv
2020-04-07 18:53:41  138978734 MDF_UNIT-run1.0-epsilon0.1-H1.csv
2020-04-07 20:44:20  139048935 MDF_UNIT-run1.0-epsilon0.5-H1.csv
2020-04-07 23:08:48  139055187 MDF_UNIT-run1.0-epsilon1.0-H1.csv
.
.
.
```

In addition, the DAS team's prepared a number of visualizations of errors on these runs. They're in this folder on DEMO_SHARE:

I don't think most of the folks on this thread have access to that folder, but I think Matt is seeking to gain access. Others might do so as well, if Tori and John are OK with it. The plots depict L1 error for each geographic unit, averaged over trials, with observations separated into bins based on the reference tabulation's true value in the CEF.

Best,

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, March 25, 2020 10:14 AM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: data from recent Virginia runs

And changing thread title.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, March 25, 2020 10:14 AM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Bumping for Jason, and adding Cynthia.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Sent: Wednesday, March 25, 2020 9:41 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Phil, thanks for sharing. Heather, Ben, Luke -- I've had some experience pulling data from S3 onto science1, so let me know if you'd like to work together on this.

Matthew Spence, Branch Chief
Foreign-Born Population Branch
Population Division
U.S. Census Bureau
o: 301-763-1033
census.gov | @uscensusbureau

Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, March 25, 2020 9:07 AM
To: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

+ Matt Spence, Jason D (will need to share the same information with them anyway!). Feel free to add anyone else you want to, to this thread (people who need/have access to science1 and will interact with the data directly).

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, March 25, 2020 8:52 AM
To: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Hi Heather/all,

Feel free to cc me when you write to Christina.

re: file locations to get started -- we're currently generating Persons-universe VA-level data, 10 repetitions, for a single PLB (4.0, the same one used for the Persons in the CNSTAT DDP release), at the same schema/scale as the CNSTAT DDP release, for a baseline (corresponding to the CNSTAT DDP setting) as well as a number of algorithms we expect to help control the positive bias problem.

This data was output to these s3 locations:

abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/heiss002/ | grep cnstat.*version2

PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries1_version2/
PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries2_version2/
PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries3_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_cnstatGeolevels_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_dpQueries1_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_dpQueries2_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_dpQueries3_version2/
PRE cnstatDdpSchema_TwoPassBigSmall_va_dpQueries1_version2/
PRE cnstatDdpSchema_TwoPassBigSmall_va_dpQueries2_version2/
PRE cnstatDdpSchema_TwoPassBigSmall_va_dpQueries3_version2/
PRE cnstatDdpSchema_nodetail_va_dpQueries1_version2/
PRE cnstatDdpSchema_nodetail_va_dpQueries2_version2/
PRE cnstatDdpSchema_nodetail_va_dpQueries3_version2/
PRE cnstatDdpSchema_nodetailsmall_va_dpQueries1_version2/
PRE cnstatDdpSchema_nodetailsmall_va_dpQueries2_version2/
PRE cnstatDdpSchema_nodetailsmall_va_dpQueries3_version2/

Notes:

- 'SinglePassRegular' denotes our current/original algorithm. The cnstatDpqueries_cnstatGeolevels variant, specifically, corresponds to what we used to generate the CNSTAT DDP publicly released data. All other output differs from that release in geohierarchy used, DP measurements taken (with 3 types considered), or algorithm used (e.g., TwoPassBigSmall, DataInd..., nodetail..., etc)
- *TwoPassBigSmall* failed on dpQueries2 & dpQueries3 unexpectedly, so any data in that location is partial
- All other runs seem to have completed successfully
- Pipe-delimited microdata versions of the output data, with header & some metadata, is located at locations like: abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/heiss002/cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_cnstatGeolevels_version2/data-run | grep .*DHCP_MDF.txt

2020-03-23 15:21:08 511045818 data-run1.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 08:55:14 511045971 data-run10.0-epsilon4.0-DHCP_MDF.txt
2020-03-23 17:11:57 511045901 data-run2.0-epsilon4.0-DHCP_MDF.txt
2020-03-23 19:04:41 511045972 data-run3.0-epsilon4.0-DHCP_MDF.txt
2020-03-23 20:53:54 511045765 data-run4.0-epsilon4.0-DHCP_MDF.txt
2020-03-23 22:48:11 511045920 data-run5.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 00:44:38 511045946 data-run6.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 02:52:23 511046014 data-run7.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 05:06:30 511045825 data-run8.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 07:03:57 511045978 data-run9.0-epsilon4.0-DHCP_MDF.txt

- For VA, output data is a little under 500 MB per .txt

Further usage notes about the individual Linux machine itself, *science1*:

- *science1* runs RedHat Enterprise Linux 7.6, which is like the RHEL around the rest of the Bureau except (a lot) more recent in vintage, so it should look/behave in familiar command-line ways
- Due to weird infrastructure choices, the root volume of *science1* is not encrypted; it's also quite small, though, and we've made it difficult to get into (you shouldn't be able to do so). If you suspect you've somehow gotten into it, though, please let us know
- Data stored on the two large (each is 16 TB) attached EBS volumes (EBS is a kind of long-term storage specific to

Amazon Web Services) */data/* and */apps/* is encrypted. /apps/ is primarily used for software, although it can serve as an alternative workspace (our software doesn't take 16 TB of space!) if you can't fit your work in /data/

- By default you don't have any folders to work in where you have permissions, but I just ssh'd in and sudo'd to root, then created 3 folders & made them (more than) permissive enough that you can work in them. I didn't know your JBIDs, so I just concat'd first initial + last name:
  - [root@ir7dassv001 data]# ls /data/

bbolender burton hking home irimata lecle301 lost+found lrogers spence tmp

Further usage notes on s3:

- s3 is a large (essentially infinite, for our purposes) remote repository in which we pay Amazon a monthly fee depending on amount of data stored/number of uploads/downloads/etc
- **EXTREMELY IMPORTANT: b**y 'remote repository', I mean that s3 'space' doesn't live on science1, so you can't cd into s3 directories, or ls them, etc
- To interact with s3, you'll need to use aws s3 API commands, which generally look similar to their *bash* command-line counterparts. The most important ones are probably:
  - *aws s3 ls* (display contents of a 'directory' in s3; technically s3 doesn't have directories, just lots of individual binary files with really long names coincidentally containing lots of forward slashes, but for our convenience the s3 API displays s3 locations' contents as-if they were directories)
  - *aws s3 cp* (upload/download, although you should only have the power to download to this machine)
  - *aws s3 sync* (upload/download an entire directory, with automatic checking of whether files uploaded/downloaded have changed before bothering to upload/download)

I think that should do it. Your basic usage pattern will probably look something like:

- I decided I want to look at this file: aws s3 ls s3://uscb-decennial-ite-das/users/heiss002/cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_cnstatGeolevels_version2/data-run | grep .*DHCP_MDF.txt

2020-03-23 15:21:08 511045818 data-run1.0-epsilon4.0-DHCP_MDF.txt

- So I cd into my directory *cd /data/myName/*
- And I cp the file down to the present location, without changing its name: *aws s3 cp s3://uscb-decennial-ite-das/users/heiss002/cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_cnstatGeolevels_version2/data-run1.0-epsilon4.0-DHCP_MDF.txt .*
- Then work as you would normally (either edit the file on *science1* with available text editors / Python / R, and maybe SAS if it is functional, or if you don't find *science1* amenable and you have space/patience to wait for transfer times, you can *scp* files back to internal T13-approved environemnts. If you do decide to do so, please let us know where you plan to transfer files, so we can get those locations added to the project's approved work locations)

If you have trouble, let me know. (Probably some small things will go wrong when you first try to work; e.g., maybe the location of *AWS* won't be setup in your user's config files and we'll have to fix that, or some such.)

Best,

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Heather King (CENSUS/POP FED) <heather.king@census.gov>
Sent: Tuesday, March 24, 2020 12:55 PM

To: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Hi all,

I say we reach out to Christina and ask her if Phil can share her materials with us. **@Phil**, can you put in a good word for us? :) **@Ben**, can you do the honors and contact Christina?

Also, **@Phil**, thanks for helping us get access to science1! I was able to log in using a SSH client. Do you have any file locations within the server that can get us started. Again, much appreciated!

-Heather

Heather King, Statistician (Demography)
Population Division/Local Government Estimates and Migration Processing Branch/HQ
U.S. Census Bureau
O: 301-763-7966
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Sent: Friday, March 13, 2020 4:20 PM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Possibly, although as we move further from the idea of using the entire MDF, it starts to lose priority. That said, it seems like a promising way to implement formal privacy while avoiding some of the concerns about post-processing for reasonableness. I'm pretty sure this still ends up with bias (inverse to cell size), but I'm wondering if it reduces it.

I say we should reach out to her. More information is rarely worse.

Ben Bolender, PhD
Senior Advisor
Population Division
U.S. Census Bureau
Phone: 301-763-9733
Email: benjamin.c.bolender@census.gov

From: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Sent: Wednesday, March 11, 2020 6:58 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Hi Philip,

Ben can correct me if I'm wrong, but after reviewing her research, it looks very promising to us. However, we're still waiting on some guidance regarding what we are expected to be producing, so right now we're just discussing how her research, as we understand it, would be applied to our process. That said, it might not hurt to have some more information about her research - Ben/Heather - what do you think? Should we reach out to Christina to see if the DAS team can share her work with us?

I'm sure if we end up getting to a point where we're starting to test her method that we'd probably be interested in reaching out to her for more information and guidance on implementation.

Thanks and I hope you're having a nice morning!

Luke

Luke T. Rogers, PhD
Chief, Population Estimates Branch
Population Division
U.S. Census Bureau
O: 301-763-7147
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Tuesday, March 10, 2020 6:05 PM
To: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Think I mentioned this elsewhere, but just in case: Christina shared an incomplete draft copy of their paper (and some code) with us about 2 weeks ago, but she asked us to keep it within the DAS team.

If you all still wanted to follow up on her work, might pay off to send her a quick email inquiring whether we could share the draft/code with you all, too?

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Sent: Monday, February 10, 2020 3:31 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Thanks Phil. It's not super urgent, we are kind of just looking for a glimmer of hope here

---
Ben Bolender, PhD

Senior Advisor
Population Division
U.S. Census Bureau
Phone: 301-763-9733
Email: benjamin.c.bolender@census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, February 10, 2020 3:30 PM
To: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

re: paper -- not sure there *is* a paper, yet. If there is, I don't have a copy of it, though.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, February 10, 2020 3:30 PM
To: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

I do! About to run out the door, but will send once I get home.

We've also separately inquired about possible use of some of Christina's code for the DAS -- you might separately poke her and ask about use of it for POP estimates? Her email should be at the end of the set of slides, once I get a chance to send 'em.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Sent: Monday, February 10, 2020 2:02 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Christina Ilvento: Slides and Paper

Hey Phil,

I heard you might have copies of the slide deck and or paper for the R&M guest presentation last week. We would like to look into the method as one of the alternatives for the population estimates.

Ben Bolender, PhD
Senior Advisor
Population Division
U.S. Census Bureau
Phone: 301-763-9733
Email: benjamin.c.bolender@census.gov

EXHIBIT 26

To: Victoria Velkoff (CENSUS/ADDP FED)[Victoria.A.Velkoff@census.gov]
From: Cynthia Davis Hollingsworth (CENSUS/DCMD FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=8CB2BA79D7DD490E807842D8ECAFBC5B-HOLLINGSWOR]
Sent: Tue 7/21/2020 12:58:38 PM (UTC)
Subject: Re: Large Epsilon Additional Runs for Review

Yep. And Phil acknowledges it's in the DP measurements too (in other words, not due to post processing).

Cynthia Davis Hollingsworth
Program Manager, 2020 Census Data Products and Dissemination
Decennial Census Management Division
U.S. Census Bureau
Office: 301.763.3655
iPhone: 202.253.6334
E-mail: cynthia.davis.hollingsworth@census.gov

From: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>
Sent: Tuesday, July 21, 2020 8:04 AM
To: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Fw: Large Epsilon Additional Runs for Review

So even with an epsilon of 500 there are errors?

Victoria Velkoff, PhD
Associate Director for Demographic Programs
U.S. Census Bureau
o: 301-763-1372
Shape your future. START HERE >2020census.gov
census.gov | @uscensusbureau

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Tuesday, July 21, 2020 7:42 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

This set appears to use one of the multipass NNLS variants. Is that correct?

Yes, it uses the basic multipass NNLS (not Robert's OLS-improved variant).

Unless I misunderstand something, the algorithm is monotonic with this rounder, but even at PLB 500, we have substantial error at every geographic level. That can't be due to the DP measurements, which all have standard deviations < 1, so it has to be due to NNLS.

The budget split for this run was:

```
epsilon_budget_total=
%(epsilon)s
                    geolevel_budget_prop= 0.2, 0.2, 0.15, 0.15, 0.15, 0.15
                    detailedprop= 0.1
                    dpqueries= total, hhgq * votingage * numraces, hhgq * votingage * hispanic * numraces, hhgq * votingage *
                    hispanic * cenrace
                    queriesprop= 0.3, 0.15, 0.15, 0.3
                    L2_DPqueryPart0= total
                    L2_DPqueryPart1= hhgq * votingage * numraces
                    L2_DPquerypart2= hhgq * votingage * hispanic * numraces
                    L2_DPqueryPart3= hhgq * votingage * hispanic * cenrace
                    L2_DPqueryPart4= detailed
```

(https://github.ti.census.gov/CB-DAS/das_decennial/blob/b7625df532b6567c3166e7994c8dd3b3e8d21da1/configs/full_person/multiL2_singlePassRounder_RI_nested.ini#L73)

At epsilon of 500, this leaves the detailed query (as the worst-case example) with "local" budget of

```
>>> 500. * 0.1 * 0.15
7.5
```

That is expended on ~500K samples per geographic unit, which still contain substantial error, even just in the DP measurements:

```
>>> x, y = prng.geometric(p, 500000) - 1., prng.geometric(p, 500000) - 1.
>>> np.sum(np.abs(x - y))
23458.0
```

That is not true for some of the other statistics in use here. For example, hhgq * votingage * hispanic * cenrace has "local" epsilon:

```
>>> epsilon = 500. * 0.3 * 0.15
>>> epsilon
22.5
```

Which will typically have 0 error in a single geounit:

```
>>> x, y = prng.geometric(p, 8*2*2*63) - 1., prng.geometric(p, 8*2*2*63) - 1.
>>> np.sum(np.abs(x - y))
0.0
```

Though this isn't true at lower geographic levels. RI has ~25K Census blocks, for example, so at that local level error in DP measurements is still non-zero, despite the larger budget assigned to this query:

```
>>> x, y = prng.geometric(p, 8*2*2*63*25000) - 1., prng.geometric(p, 8*2*2*63*25000) - 1.
>>> np.sum(np.abs(x - y))
1267.0
```

That said, I suspect the detailed query may be inflating the error somewhat. Since it works over the floats, not the integers,

multipass doesn't constrain later passes to exactly match query estimates from earlier passes; instead, it only requires they agree within a tolerance, which I have set to 5.0 by default:

https://github.ti.census.gov/CB-DAS/das_decennial/blob/29a2823973d90c4b724dc437c18620cff7fc4f83/programs/optimization/l2_dataIndep_npass_optimizer.py#L144
https://github.ti.census.gov/CB-DAS/das_decennial/blob/29a2823973d90c4b724dc437c18620cff7fc4f83/programs/optimization/l2_dataIndep_npass_optimizer.py#L176

As a result, the larger error in the detailed query can mildly deteriorate the other estimates in each geounit. And this effect can compound additively as we move down the geohierarchy -- introducing 5 extra error at the County level, then 5 more at the Tract_Group level, and so on (and this can happen within each scalar of each vectorized query, so it is worse for the 8*2*2*63 query than for the Total Pop query).

A few other things to note:

- multipass is likely to converge more slowly to 0 error than our original approach with a single simultaneous optimization (though Robert's OLS-improved multipass should help with this)
- there is room to modify the budget settings; this is still quite a preliminary run. We might try re-allocating some PLB away from total pop, and away from the top two geolevels
- I could try reducing the multipass tolerance a bit, although this will eventually induce instability
- currently, the NNLS solve and the Rounder target the same queries, and we recently learned that this new Rounder requires "hierarchically nested" queries, so I was forced to modify the NNLS queries to respect this property as well. This restriction can be lifted, though, which would allow us more freedom in the NNLS query specification (which may be to the benefit of convergence rate for our queries of interest)

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Monday, July 20, 2020 11:27 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Easiest way is to just turn on the viewer and rummage around.

Unless I misunderstand something, the algorithm is monotonic with this rounder, but even at PLB 500, we have substantial error at every geographic level. That can't be due to the DP measurements, which all have standard deviations < 1, so it has to be due to NNLS. This set appears to use one of the multipass NNLS variants. Is that correct?

Crashed my Chrome, so I invoked my "first crash after 9pm = all done" rule. All done for today.

Thanks,

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

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, July 20, 2020 3:07 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

If you mean our (DAS team analysis module) usual error visualizations, John, I've been doing so; the script did not like my asking it to process a much larger number of parameter variations for usual, so I'm having to fill in combinations it missed somewhat piecemeal, but I've been documenting this in:

https://github.ti.census.gov/CB-DAS/das_decennial/issues/393#issuecomment-12534

The primary folder of interest is:

RM_SHARED:\dms-p0-992\CNSTAT_DDP_Improvement_Experiments_March_2020\asymptotic_epsilon_investigation\lecle301\nested_queries_multiL2_singlePassRounder\visualizations\singlePassRounder_VA

Images are best opened [A] as pngs and [B] in Chrome (or the point at which the volume is mounted moved down several sub-folders) (otherwise, Windows will complain about too-long path names)

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Monday, July 20, 2020 3:02 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Thanks. Could you please stage the salient ones to dms-p0-992?

John M. Abowd, PhD, Associate Director and Chief Scientist
Research and Methodology
U.S. Census Bureau
O: 301-763-5880 M: simulring on cell
census.gov | @uscensusbureau

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, July 20, 2020 2:23 PM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Cc: Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Hi Matt (+ Tori, John, Cynthia to cc),

We've made some notable progress on the large-epsilon runs, though we haven't tried any Nation-wide yet and we have some further tuning of budget parameters to do, and one more refinement of the method to implement.

If you (or others with access to/knowledge of the metrics scripts) have time, you may want to analyze the following VA and RI data (the bolded, underlined ones):

tacos-ITE-MASTER:hadoop@ip-10-252-47-18$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_multiL2_

PRE cnstatDdpSchema_multiL2_cellWiseRounder_nested_accuracyTest/
PRE cnstatDdpSchema_multiL2_cellWiseRounder_nested_accuracyTest_RI/
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest/
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest_-1./
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest_100/
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest_eps100_singleRounderPass/
PRE cnstatDdpSchema_multiL2_multiRounder_nonmonotonicityTest_manyEps_singleRounderPass/
PRE cnstatDdpSchema_multiL2_singlePassRounder_nested_accuracyTest/
PRE cnstatDdpSchema_multiL2_singlePassRounder_nested_accuracyTest_RI/

Of the four, the two marked RI are, well, RI data, and the two not marked RI are (poorly named) VA data.

cellWiseRounder denotes our original rounder. singlePassRounder is one of the two new methods, which from what I can tell works as intended (largely -- maybe completely -- restoring monotonicity of accuracy in epsilon); it adds L1 penalty terms to the Rounder, in addition to the Rounder's usual detailed-cell terms.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, June 1, 2020 1:09 PM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Hey Matt,

Sort of -- we did *many* re-runs, and conducted a lengthy investigation (mostly documented here, if you're curious: https://github.ti.census.gov/CB-DAS/das_decennial/issues/264), but none of them would be useful for you all to analyze

right now. Basically, we identified two issues that need to be fixed before it is probably worthwhile to re-run large *epsilon* runs for proper Analysis:

- a while back, to improve the security of our pseudo-random number generator, we switched from using standard *numpy* random distributions/samplers to using an Intel distribution of Anaconda that has an alternative library, *mkl_random*, for doing the same thing. Unfortunately, we didn't realize that *mkl_random* makes several non-obvious, unadvertised changes -- including, notably, that it behaves improperly (gives nonsensical errors) at degenerate scale parameters (including, specifically, that it yields arbitrary, large perturbations when fed a scale value of *1.0* for a Geometric distribution; this is the opposite of the way base *numpy* operates, which behaves as you'd expect it to in the continuous limit, i.e., yields negligible/zero noise at scale 1.0)
- as *epsilon* increases, the importance of the Rounder problem (which converts float-valued estimates to integer-valued estimates) increases. This is unfortunately expensive (in terms of accuracy), because the Rounder is not designed to be as statistically efficient as the main NNLS solves (a sacrifice made because the Rounder has to be structurally simpler to guarantee that it will be able to find an integer-valued solution at all)

The fix for the first problem is pretty simple, and we can implement it and guarantee that, eventually, large epsilon will give perfect accuracy.

But that's probably not enough; the second problem requires a little more work to fix, but is very important, because it can create "non-monotonicities": queries can get worse as epsilon increases (after post-processing, because the Rounder has become important), before eventually getting better (because even the Rounder is eventually perfect, if 0 noise is introduced). That can make the relationship of *epsilon* to accuracy more complicated than we want it to be, so we need to implement improvements to tackle this second problem before we re-generate "official" large-epsilon runs.

Well, unless we want epsilon so large that accuracy is near-perfect, anyway. That can be done with just the simple fix, but I think it is probably not sufficient.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Sent: Monday, June 1, 2020 12:33 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Hi Phil-

I hope you're well. Did you ever do a re-run of these large epsilon runs?

Thanks,
-Matt

Matthew Spence, Branch Chief
Foreign-Born Population Branch
Population Division
U.S. Census Bureau
o: 301-763-1033

Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, May 18, 2020 10:03 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Large Epsilon Additional Runs for Review

Hi all,

I've completed initial Analyses on the large-epsilon runs, but the results look somewhat odd (implausibly high error). Would hold off on reviewing them until we can investigate.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Friday, May 15, 2020 10:33 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Large Epsilon Additional Runs for Review

Hi all,

After conferring with John and Dan K. on Wednesday, we elected to generate additional National runs at very large epsilons. Two new runs, using each of singlePassRegular (basic TopDown) and dataIndUserSpecifiedMultipass (basic multipass) and total *epsilon=100*, have completed and are available for analysis at:

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_National_dpQueries100_largeEpsRun/

and

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries100_largeEpsRun/

Note that, even at *epsilon=100*, non-zero noise was almost certainly introduced (because the budget is divided in 5 parts per level, for 7 levels, so the on-average local *epsilon* expenditure on a query is approximately *100/35* -- which is still large, but not so large as to expect 0 noise over billions of noisy samples).

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Saturday, May 2, 2020 10:05 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/EWD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP Wigglesum runs # 1-4 (Persons Universe) for CNSTAT Experts Meetings

The dpQueries5 wigglesum run has completed now as well. It is at:

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_TwoPassBigSmall_National_dpQueries5_AllChildFilter

(this is the same location as in the last email, but at the time the dpQueries5 s3 "directory" was only partially populated)

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Friday, May 1, 2020 3:47 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/EWD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Official CNSTAT DDP Wigglesum runs # 1-4 (Persons Universe) for CNSTAT Experts Meetings

Hi all,

Note changed thread title. The dpQueries1-4 Wigglesum runs are complete; their microdata is available in our s3 bucket

at the underlined/bolded/italicized locations below:

brach-ITE-MASTER:hadoop@ip-10-252-47-189$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_TwoPassBigSmall_National_dpQueries

PRE ***cnstatDdpSchema_TwoPassBigSmall_National_dpQueries1_AllChildFilter/***
PRE cnstatDdpSchema_TwoPassBigSmall_National_dpQueries2_AllChildFilter/
PRE ***cnstatDdpSchema_TwoPassBigSmall_National_dpQueries2_AllChildFilter_attempt2/***
PRE cnstatDdpSchema_TwoPassBigSmall_National_dpQueries3_AllChildFilter/
PRE ***cnstatDdpSchema_TwoPassBigSmall_National_dpQueries3_AllChildFilter_attempt2/***
PRE ***cnstatDdpSchema_TwoPassBigSmall_National_dpQueries4_AllChildFilter/***
PRE cnstatDdpSchema_TwoPassBigSmall_National_dpQueries5_AllChildFilter/

The dpQueries5 Wigglesum run will complete sometime tomorrow morning (5/2/2020). For dpQueries2 and dpQueries3, please use the "attempt2" locations (as bolded above); the first two attempts for those budget settings failed because we ran out of gurobi licenses. (That is also the reason for the small delay in this email, and in delivery of dpQueries5 MDF data; as it turns out, we do not have enough gurobi licenses to execute 4 National runs simultaneously!)

Best,

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Tuesday, April 28, 2020 8:17 AM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

Standard DAS error visualizations have been prepared and downloaded to the dms-p0-992 folder. Currently copying them over to DAS_Collaboration. Four new folders were added in *Y:\CNSTAT_DDP_Improvements\preliminary_analyses\Persons*

cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries2_officialCNSTATrun
cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries3_officialCNSTATrun
cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries4_officialCNSTATrun
cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries5_officialCNSTATrun

As previously indicated, these differ from one another in that, although all have the same overall privacy-loss budget of 4.0, dpQueriesX assigned increasing proportions of the privacy-loss budget to the Total query as X gets larger. This should improve error on Total Population, but worsen error on other tabulations.

These should be compared with the original CNSTAT run (noting that some of the labels were out-of-order in the original National CNSTAT run Analysis), in sub-folder:

cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries1_officialCNSTATrun

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, April 27, 2020 3:23 PM
To: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

Update: looks like the visualizations won't be ready this evening; all 4 analysis scripts are still running, but are in the middle of the PL94 tabulations (perhaps I should have used larger clusters).

I'll check on them again later tonight, but it seems likely they may have to run overnight.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Sent: Monday, April 27, 2020 10:35 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

Got it. Thanks. Planning to review this evening if the visualizations are ready.

John M. Abowd, PhD, Associate Director and Chief Scientist
Research and Methodology
U.S. Census Bureau
O: 301-763-5880 M: simulring on cell

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, April 27, 2020 10:33 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

p.s. Oops, I left a typo here -- this should say dpQueries***1***:

dpQueries2, 0.05 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries1_officialCNSTATrun.ini#L63 (this run is not new; it was the original Persons universe run, which we examined Matt's error metrics on last Wednesday)

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, April 27, 2020 10:32 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 2-5 (Persons Universe) for CNSTAT Experts Meetings

Hi all,

Note changed thread title.

After review of error metrics for the # 1 CNSTAT Experts meeting run for the Persons universe last Wednesday, I was asked to generate additional runs, with increasing proportions of the overall privacy-loss budget [PLB] assigned to the Total Population query. I believe John and Tori will select one of these runs to use as the 'official Persons-universe run (or, if none of these seem acceptable, may ask for additional run(s)). These runs are complete, and can be found at:

abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQuerie
PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries1_officialCNSTATrun/
PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries2_officialCNSTATrun/
PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries3_officialCNSTATrun/

PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries4_officialCNSTATrun/
PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries5_officialCNSTATrun/

The budget sections of the configuration files show the PLB allocation and measurements taken for each of these runs:

dpQueries2, 0.05 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries1_officialCNSTATrun.ini#L63 (this run is not new; it was the original Persons universe run, which we examined Matt's error metrics on last Wednesday)

dpQueries2, 0.2 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/328209709cac43fbaf2c1062d8271812269056a3/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries2_officialCNSTATrun.ini#L63

dpQueries3, 0.3 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries3_officialCNSTATrun.ini#L63

dpQueries4, 0.5 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries4_officialCNSTATrun.ini#L63

dpQueries5, 0.75 proportion to Total : https://github.ti.census.gov/CB-DAS/das_decennial/blob/master/configs/full_person/DAS_NAT_DHCP_HHGQ_NNLS_vs_OLS_Experiment_dataIndMultipass_dpQueries5_officialCNSTATrun.ini#L63

Note that, as the proportion of PLB assigned to Total increases, we expect error generally to increase on the remaining queries/tabulations.

I am currently working on the standard DAS Analyses visualizations; they're not ready yet, but I will copy them to DAS Collaboration and update here when they are prepared (likely by COB today).

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, April 22, 2020 9:49 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Re: Official CNSTAT DDP (re-)runs # 1 for CNSTAT Experts Meetings

Hi all,

I was asked in our 9 AM meeting today when the official Nation-wide CNSTAT Experts runs, & unpickled H1-only measurements, were delivered.

Please see the below email from 4/15/2020; it constituted that delivery.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, April 15, 2020 2:13 PM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>; Simson L Garfinkel (CENSUS/ADRM FED) <simson.l.garfinkel@census.gov>
Subject: Official CNSTAT DDP (re-)runs # 1 for CNSTAT Experts Meetings

Hi all (also, + Simson to cc),

Note re-named thread title.

The Units (H1-only) and Persons National runs (using the `DHCP_HHGQ` schema, which is the same histogram we relied on for the CNSTAT DDP release) for the next CNSTAT Experts meeting have both completed, and individual algorithm variants/trials have been selected by Tori and John as those we'll use for the next experts update. In addition, I've fixed the header de-duplication issue in the H1-only measurements, and unpickled the H1-only measurements for the official Units run.

The official data locations in s3 are as follows.

Persons MDF:

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_National_dpQueries1_officialCNSTATrun/data-run1.0-epsilon4.0-DHCP_MDF.txt

H1-only Units MDF:

s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_nat_H1_Only_withMeasurements_v8/MDF_UNIT-run1.0-epsilon0.5-H1.csv

H1-only unpickled pipe-delimited measurements text files (Occ/Vac count estimates, integer but can be negative; consistent, because there were only the two values measured; one file per geolevel; each file with header)**:**

Within s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_nat_H1_Only_withMeasurements_v8/noisy_measurements-eps0.5-run1/

The relevant files are:

application_1586267880679_0010-Block.csv
application_1586267880679_0010-Block_Group.csv
application_1586267880679_0010-Tract.csv
application_1586267880679_0010-Tract_Group.csv
application_1586267880679_0010-County.csv
application_1586267880679_0010-State.csv
application_1586267880679_0010-National.csv

I think this concludes this first pass-off of data from the DAS team to POP/DEMO for the experts meetings. Please let me know if there are questions/concerns/access issues, etc.

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Thursday, April 9, 2020 7:07 PM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Victoria Velkoff (CENSUS/ADDP FED) <Victoria.A.Velkoff@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Christine Flanagan Borman (CENSUS/POP FED) <christine.flanagan.borman@census.gov>
Cc: John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>
Subject: Re: data from recent Virginia runs

Hi everyone,

Adding Jason, Tori, John, Christine B. to this thread. (Please add anyone else who should have access to this data.)

In addition to sharing experimental MDFs (post-processed DP measurements) from a variety of algorithms, the DAS team was asked/had promised to provide DP measurements (before post-processing) for one of the H1-only runs. This will be a good starting point for understanding them, because the H1-only measurements are very simple (just two numbers per geounit). For POP/DEMO folks with access to our s3 bucket, I've provided this data for each geographic level in the first PLB 0.007 H1-only run, here:

abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_nat_H1_Only_withMeasurements_v8/noisy_measurements-eps0.007-run1/application_158626788 | grep .*\.csv
PRE application_1586267880679_0010-Block.csv/

PRE application_1586267880679_0010-Block_Group.csv/
PRE application_1586267880679_0010-County.csv/
PRE application_1586267880679_0010-National.csv/
PRE application_1586267880679_0010-State.csv/
PRE application_1586267880679_0010-Tract.csv/
PRE application_1586267880679_0010-Tract_Group.csv/

2020-04-09 18:36:42 application_1586267880679_0010-Block.csv
2020-04-09 18:39:20 application_1586267880679_0010-Block_Group.csv
2020-04-09 18:50:13 application_1586267880679_0010-County.csv
2020-04-09 18:55:08 application_1586267880679_0010-National.csv
2020-04-09 18:51:23 application_1586267880679_0010-State.csv
2020-04-09 18:43:26 application_1586267880679_0010-Tract.csv
2020-04-09 18:48:38 application_1586267880679_0010-Tract_Group.csv

I think the measurement files' structure should be self-explanatory, but let me know if there are questions. Also please let me know if you think you've found any errors in them, too, of course!

Two important notes:

- an unfortunate quirk remains -- I accidentally let the concatenation program duplicate the csv's header many times. So, when parsing these, you'll have to remove rows that look like: *DPQueryName|geocode|Vacant|Occupied*. There will be more than just one such row! (I can fix this early next week, but this caveat applies for anyone who uses this data before then.)
- the PLB in this case is *very* small, but it is also what I crudely estimate we expended on estimating the number of Occupied Housing Units in the CNSTAT DDP release (in case you were wondering where 0.007 came from). We also have runs for PLBs 0.1, 0.5, 2.0, 1.0, and 4.0, and I could similarly unwrap DP measurements for some of those runs, if there's interest

Also, for H1-only, these files are relatively small, so moving them to other secure locations should be relatively easy, if desired. (That will not be the case when we share the DP measurements for full runs; for DHC(-P), for example, the full measurement sets tend to be measurements in dozens of terabytes.)

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Thursday, April 9, 2020 4:57 PM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Hi all,

I've corrected the incorrect visualizations and replaced the copy in DAS_Collaboration.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Thursday, April 9, 2020 11:41 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Oh, an unfortunate update on the visualizations in *preliminary_analyses*: in the most recent runs, they computed *average signed error*, not *mean absolute error*, because of a change I made to improve scalability (not noticing that I had lost a *.abs* in the process).

They'll have to be re-computed with the *.abs* fixed.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Sent: Thursday, April 9, 2020 8:28 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Thanks Matt and Philip! I'll submit the DIRT ticket this morning.

Luke

Luke T. Rogers, PhD
Chief, Population Estimates Branch
Population Division
U.S. Census Bureau
O: 301-763-7147
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Thursday, April 9, 2020 8:24 AM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

That's right, Matt, yep.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Sent: Thursday, April 9, 2020 8:04 AM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Hi Luke -

I did complete a DIRT ticket request for \\it171oafs-oa03.boc.ad.census.gov\DEMO_SHARE\DAS Collaboration yesterday afternoon. I believe \CNSTAT_DDP_Improvement_Experiments\preliminary_analyses\ is nested within DAS Collaboration. (Phil, is that right?)

Matthew Spence, Branch Chief
Foreign-Born Population Branch
Population Division
U.S. Census Bureau
o: 301-763-1033
census.gov | @uscensusbureau

Shape your future. START HERE > 2020census.gov

From: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Sent: Thursday, April 9, 2020 6:40 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Thank you Philip! Excited to see your team's results.

I would like access to \CNSTAT_DDP_Improvement_Experiments\preliminary_analyses\ .

Matt - did you just request access through DIRT?

Luke

Luke T. Rogers, PhD
Chief, Population Estimates Branch
Population Division
U.S. Census Bureau
O: 301-763-7147
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, April 8, 2020 5:30 PM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: data from recent Virginia runs

Hi all,

We have more data, now. For the Persons universe, the version7 and v7 runs here are good for analysis:

```
abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_ | grep [4,6]_v.*7
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries4_version7/
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries6_v7/
```

We also have Housing Unit (just Occupied/Vacant, per previous conversations) runs at various epsilons. They can be found here:

```
hazan-ITE-MASTER:hadoop@ip-10-252-46-63$ aws s3 ls s3://uscb-decennial-ite-das/users/lecle301/cnstatDdpSchema_SinglePassRegular_na_H1_Only_withMeasurements_v8/ | grep .*\.csv

2020-04-07 16:18:23  138923937 MDF_UNIT-run1.0-epsilon0.007-H1.csv
2020-04-07 18:53:41  138978734 MDF_UNIT-run1.0-epsilon0.1-H1.csv
2020-04-07 20:44:20  139048935 MDF_UNIT-run1.0-epsilon0.5-H1.csv
2020-04-07 23:08:48  139055187 MDF_UNIT-run1.0-epsilon1.0-H1.csv
.
.
.
```

In addition, the DAS team's prepared a number of visualizations of errors on these runs. They're in this folder on DEMO_SHARE:

DEMO_SHARE:\CNSTAT_DDP_Improvement_Experiments\preliminary_analyses\

I don't think most of the folks on this thread have access to that folder, but I think Matt is seeking to gain access. Others might do so as well, if Tori and John are OK with it. The plots depict L1 error for each geographic unit, averaged over trials, with observations separated into bins based on the reference tabulation's true value in the CEF.

Best,

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, March 25, 2020 10:14 AM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: data from recent Virginia runs

And changing thread title.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, March 25, 2020 10:14 AM
To: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Bumping for Jason, and adding Cynthia.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>
Sent: Wednesday, March 25, 2020 9:41 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Phil, thanks for sharing. Heather, Ben, Luke – I've had some experience pulling data from S3 onto science1, so let me know if you'd like to work together on this.

Matthew Spence, Branch Chief
Foreign-Born Population Branch
Population Division
U.S. Census Bureau
o: 301-763-1033
census.gov | @uscensusbureau

Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, March 25, 2020 9:07 AM
To: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Matthew Spence (CENSUS/POP FED) <Matthew.Spence@census.gov>; Jason Devine (CENSUS/POP FED) <Jason.E.Devine@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

+ Matt Spence, Jason D (will need to share the same information with them anyway!). Feel free to add anyone else you want to, to this thread (people who need/have access to science1 and will interact with the data directly).

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Wednesday, March 25, 2020 8:52 AM
To: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Hi Heather/all,

Feel free to cc me when you write to Christina.

re: file locations to get started -- we're currently generating Persons-universe VA-level data, 10 repetitions, for a single PLB (4.0, the same one used for the Persons in the CNSTAT DDP release), at the same schema/scale as the CNSTAT DDP release, for a baseline (corresponding to the CNSTAT DDP setting) as well as a number of algorithms we expect to help control the positive bias problem.

This data was output to these s3 locations:

```
abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/heiss002/ | grep cnstat.*version2
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries1_version2/
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries2_version2/
                           PRE cnstatDdpSchema_DataIndUserSpecifiedQueriesNPass_va_dpQueries3_version2/
```

PRE cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_cnstatGeolevels_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_dpQueries1_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_dpQueries2_version2/
PRE cnstatDdpSchema_SinglePassRegular_va_dpQueries3_version2/
PRE cnstatDdpSchema_TwoPassBigSmall_va_dpQueries1_version2/
PRE cnstatDdpSchema_TwoPassBigSmall_va_dpQueries2_version2/
PRE cnstatDdpSchema_TwoPassBigSmall_va_dpQueries3_version2/
PRE cnstatDdpSchema_nodetail_va_dpQueries1_version2/
PRE cnstatDdpSchema_nodetail_va_dpQueries2_version2/
PRE cnstatDdpSchema_nodetail_va_dpQueries3_version2/
PRE cnstatDdpSchema_nodetailsmall_va_dpQueries1_version2/
PRE cnstatDdpSchema_nodetailsmall_va_dpQueries2_version2/
PRE cnstatDdpSchema_nodetailsmall_va_dpQueries3_version2/

Notes:

- 'SinglePassRegular' denotes our current/original algorithm. The cnstatDpqueries_cnstatGeolevels variant, specifically, corresponds to what we used to generate the CNSTAT DDP publicly released data. All other output differs from that release in geohierarchy used, DP measurements taken (with 3 types considered), or algorithm used (e.g., TwoPassBigSmall, DataInd..., nodetail..., etc)
- *TwoPassBigSmall* failed on dpQueries2 & dpQueries3 unexpectedly, so any data in that location is partial
- All other runs seem to have completed successfully
- Pipe-delimited microdata versions of the output data, with header & some metadata, is located at locations like: abdat-ITE-MASTER:hadoop@ip-10-252-44-211$ aws s3 ls s3://uscb-decennial-ite-das/users/heiss002/cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_cnstatGeolevels_version2/data-run | grep .*DHCP_MDF.txt

2020-03-23 15:21:08 511045818 data-run1.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 08:55:14 511045971 data-run10.0-epsilon4.0-DHCP_MDF.txt
2020-03-23 17:11:57 511045901 data-run2.0-epsilon4.0-DHCP_MDF.txt
2020-03-23 19:04:41 511045972 data-run3.0-epsilon4.0-DHCP_MDF.txt
2020-03-23 20:53:54 511045765 data-run4.0-epsilon4.0-DHCP_MDF.txt
2020-03-23 22:48:11 511045920 data-run5.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 00:44:38 511045946 data-run6.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 02:52:23 511046014 data-run7.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 05:06:30 511045825 data-run8.0-epsilon4.0-DHCP_MDF.txt
2020-03-24 07:03:57 511045978 data-run9.0-epsilon4.0-DHCP_MDF.txt

- For VA, output data is a little under 500 MB per .txt

Further usage notes about the individual Linux machine itself, *science1*:

- *science1* runs RedHat Enterprise Linux 7.6, which is like the RHEL around the rest of the Bureau except (a lot) more recent in vintage, so it should look/behave in familiar command-line ways
- Due to weird infrastructure choices, the root volume of *science1* is not encrypted; it's also quite small, though, and we've made it difficult to get into (you shouldn't be able to do so). If you suspect you've somehow gotten into it, though, please let us know
- Data stored on the two large (each is 16 TB) attached EBS volumes (EBS is a kind of long-term storage specific to Amazon Web Services) */data/* and */apps/* is encrypted. /apps/ is primarily used for software, although it can serve as an alternative workspace (our software doesn't take 16 TB of space!) if you can't fit your work in /data/
- By default you don't have any folders to work in where you have permissions, but I just ssh'd in and sudo'd to root, then created 3 folders & made them (more than) permissive enough that you can work in them. I didn't know your JBIDs, so I just concat'd first initial + last name:
    - [root@ir7dassv001 data]# ls /data/

Further usage notes on s3:

- s3 is a large (essentially infinite, for our purposes) remote repository in which we pay Amazon a monthly fee depending on amount of data stored/number of uploads/downloads/etc
- **EXTREMELY IMPORTANT: b**y 'remote repository', I mean that s3 'space' doesn't live on science1, so you can't cd into s3 directories, or ls them, etc
- To interact with s3, you'll need to use aws s3 API commands, which generally look similar to their *bash* command-line counterparts. The most important ones are probably:
  - *aws s3 ls* (display contents of a 'directory' in s3; technically s3 doesn't have directories, just lots of individual binary files with really long names coincidentally containing lots of forward slashes, but for our convenience the s3 API displays s3 locations' contents as-if they were directories)
  - *aws s3 cp* (upload/download, although you should only have the power to download to this machine)
  - *aws s3 sync* (upload/download an entire directory, with automatic checking of whether files uploaded/downloaded have changed before bothering to upload/download)

I think that should do it. Your basic usage pattern will probably look something like:

- I decided I want to look at this file: aws s3 ls s3://uscb-decennial-ite-das/users/heiss002/cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_cnstatGeolevels_version2/data-run | grep .*DHCP_MDF.txt

2020-03-23 15:21:08 511045818 data-run1.0-epsilon4.0-DHCP_MDF.txt

- So I cd into my directory *cd /data/myName/*
- And I cp the file down to the present location, without changing its name: *aws s3 cp s3://uscb-decennial-ite-das/users/heiss002/cnstatDdpSchema_SinglePassRegular_va_cnstatDpqueries_cnstatGeolevels_version2/data-run1.0-epsilon4.0-DHCP_MDF.txt .*
- Then work as you would normally (either edit the file on *science1* with available text editors / Python / R, and maybe SAS if it is functional, or if you don't find *science1* amenable and you have space/patience to wait for transfer times, you can *scp* files back to internal T13-approved environemnts. If you do decide to do so, please let us know where you plan to transfer files, so we can get those locations added to the project's approved work locations)

If you have trouble, let me know. (Probably some small things will go wrong when you first try to work; e.g., maybe the location of *AWS* won't be setup in your user's config files and we'll have to fix that, or some such.)

Best,
Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Heather King (CENSUS/POP FED) <heather.king@census.gov>
Sent: Tuesday, March 24, 2020 12:55 PM
To: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Hi all,

I say we reach out to Christina and ask her if Phil can share her materials with us. **@Phil**, can you put in a good word for us? :) **@Ben**, can you do the honors and contact Christina?

Also, **@Phil**, thanks for helping us get access to science1! I was able to log in using a SSH client. Do you have any file locations within the server that can get us started. Again, much appreciated!

-Heather

Heather King, Statistician (Demography)
Population Division/Local Government Estimates and Migration Processing Branch/HQ
U.S. Census Bureau
O: 301-763-7966
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Sent: Friday, March 13, 2020 4:20 PM
To: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>; Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Possibly, although as we move further from the idea of using the entire MDF, it starts to lose priority. That said, it seems like a promising way to implement formal privacy while avoiding some of the concerns about post-processing for reasonableness. I'm pretty sure this still ends up with bias (inverse to cell size), but I'm wondering if it reduces it.

I say we should reach out to her. More information is rarely worse.

Ben Bolender, PhD
Senior Advisor
Population Division
U.S. Census Bureau
Phone: 301-763-9733
Email: benjamin.c.bolender@census.gov

From: Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Sent: Wednesday, March 11, 2020 6:58 AM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>; Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Hi Philip,

Ben can correct me if I'm wrong, but after reviewing her research, it looks very promising to us. However, we're still waiting on some guidance regarding what we are expected to be producing, so right now we're just discussing how her research, as we understand it, would be applied to our process. That said, it might not hurt to have some more information about her research - Ben/Heather - what do you think? Should we reach out to Christina to see if the DAS team can share her work with us?

I'm sure if we end up getting to a point where we're starting to test her method that we'd probably be interested in reaching out to

her for more information and guidance on implementation.

Thanks and I hope you're having a nice morning!

Luke

Luke T. Rogers, PhD
Chief, Population Estimates Branch
Population Division
U.S. Census Bureau
O: 301-763-7147
census.gov | @uscensusbureau
Shape your future. START HERE > 2020census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Tuesday, March 10, 2020 6:05 PM
To: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Think I mentioned this elsewhere, but just in case: Christina shared an incomplete draft copy of their paper (and some code) with us about 2 weeks ago, but she asked us to keep it within the DAS team.

If you all still wanted to follow up on her work, might pay off to send her a quick email inquiring whether we could share the draft/code with you all, too?

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Sent: Monday, February 10, 2020 3:31 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

Thanks Phil. It's not super urgent, we are kind of just looking for a glimmer of hope here

Ben Bolender, PhD
Senior Advisor
Population Division
U.S. Census Bureau
Phone: 301-763-9733
Email: benjamin.c.bolender@census.gov

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>

Sent: Monday, February 10, 2020 3:30 PM
To: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

re: paper -- not sure there *is* a paper, yet. If there is, I don't have a copy of it, though.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Sent: Monday, February 10, 2020 3:30 PM
To: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Re: Christina Ilvento: Slides and Paper

I do! About to run out the door, but will send once I get home.

We've also separately inquired about possible use of some of Christina's code for the DAS -- you might separately poke her and ask about use of it for POP estimates? Her email should be at the end of the set of slides, once I get a chance to send 'em.

Philip Leclerc
Mathematical Statistician
Center for Disclosure Avoidance Research (former)
Center for Enterprise Dissemination (current)
United States Census Bureau
Work Phone: 301-763-3716
Cell Phone: 202-510-0188

From: Benjamin C Bolender (CENSUS/POP FED) <Benjamin.C.Bolender@census.gov>
Sent: Monday, February 10, 2020 2:02 PM
To: Philip Leclerc (CENSUS/CDAR FED) <philip.leclerc@census.gov>
Cc: Heather King (CENSUS/POP FED) <heather.king@census.gov>; Luke Rogers (CENSUS/POP FED) <luke.rogers@census.gov>
Subject: Christina Ilvento: Slides and Paper

Hey Phil,

I heard you might have copies of the slide deck and or paper for the R&M guest presentation last week. We would like to look into the method as one of the alternatives for the population estimates.

Ben Bolender, PhD
Senior Advisor
Population Division
U.S. Census Bureau
Phone: 301-763-9733
Email: benjamin.c.bolender@census.gov

EXHIBIT 27

To: John Maron Abowd (CENSUS/ADRM FED)[john.maron.abowd@census.gov]; Robert Sienkiewicz (CENSUS/CED FED)[robert.sienkiewicz@census.gov]
From: Michael B Hawes (CENSUS/CED FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=9E817019D3624EC9B204F0DB1EE148BE-HAWES, MICH]
Sent: Fri 7/24/2020 12:40:54 PM (UTC)
Subject: Fw: fact-checking request
Goroff-Groshen for HSDR_v1.docx

John,

Given your reaction to the modernization paper, you may want to take a look at the attached draft of the Goroff-Groshen article for HDSR which they asked Rob to fact-check. It has numerous references to the modernization paper - including a reference to a 2000 agreement with DOJ about keeping block-level pop totals invariant (see page 15).

Michael B. Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology
U.S. Census Bureau
301.763.1960 (office)
202.669.9035 (mobile)
michael.b.hawes@census.gov

From: Robert Sienkiewicz (CENSUS/CED FED) <robert.sienkiewicz@census.gov>
Sent: Thursday, July 23, 2020 10:20 AM
To: Michael B Hawes (CENSUS/CED FED) <michael.b.hawes@census.gov>
Subject: Fw: fact-checking request

Hi!
Thanks for your review. I'll be looking at it as well.
Rob

Robert T. Sienkiewicz, Ph.D., MBA
Chief, Center for Enterprise Dissemination
U.S. Census Bureau
phone: 301-763-1234 (direct); 202-604-6967 (mobile)

From: Erica Groshen <erica.groshen@gmail.com>
Sent: Sunday, July 19, 2020 6:59 PM
To: Robert Sienkiewicz (CENSUS/CED FED) <robert.sienkiewicz@census.gov>
Subject: fact-checking request

Hi Robert:
Danny Goroff and I will be submitting a draft of this paper to the HSDR for the differential privacy conference volume in about 10 days. It is still a bit rough, but I want to be sure that our facts are correct now, so that I can amend anything accordingly. Could you please have someone take a look to ensure that we do not have any mistakes or facts that need updating in it? Other comments are certainly welcome, too.
I would really appreciate the help!
Thanks,
Erica
--
Erica L. Groshen
Cornell University--ILR School

EXHIBIT 29

To: James Whitehorne (CENSUS/ADDC FED)[James.Whitehorne@census.gov]; Cynthia Davis Hollingsworth (CENSUS/DCMD FED)[cynthia.davis.hollingsworth@census.gov]; Kathleen M Styles (CENSUS/ADDC FED)[kathleen.m.styles@census.gov]
Cc: Deborah Stempowski (CENSUS/ADDC FED)[Deborah.M.Stempowski@census.gov]
From: Jane H Ingold (CENSUS/DCMD FED)[/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=99DB9B5B37F44C58825FF5DB7FDFD4AD-INGOLD, JAN]
Sent: Thur 7/9/2020 3:03:53 PM (UTC)
Subject: Re: Silver Lining decades ago
Uses-of-Census-Bureau-Data-in-Federal-Funds-Distribution.pdf

We conduct the Census for reasons (mandated and required). I attached the $ depending on the Census (but not our submission to Congress document). The 2020 Census is not a science project or computer stop-gap for DAS. As of today without the awareness of many external data users:

- *noise* will be added to the undercount (PES estimation)
- check box data for race and Hispanic origin are delayed
- CQR cannot figure out what to put in a stakeholder guide or its operational bounds

---

Jane Ingold, Special Assistant,
Decennial Census Management Division, U.S. Census Bureau
Office 301.763.4646 2H172A Cell 301.775.7515
jane.h.ingold@census.gov
census.gov Connect with us on Social Media

From: James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>
Sent: Thursday, July 9, 2020 10:54 AM
To: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>; Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>; Kathleen M Styles (CENSUS/ADDC FED) <kathleen.m.styles@census.gov>
Cc: Deborah Stempowski (CENSUS/ADDC FED) <Deborah.M.Stempowski@census.gov>
Subject: Re: Silver Lining decades ago

Thanks for the background Cynthia -
I hesitated to say more in that meeting because the DAS consistently pushes against the redistricting data set as the cause of all evil in regards to accuracy of the data. They appear to use these other drastic proposals to push back against what is the well explained, documented, and expected use case for the redistricting data at the block level of geography. Based on my interactions with the DAS project over three years, even if we changed P.L. to tract level, there would be some other data or geography in the way of making their system work. The reality is that there are several use cases for small area geography for which the census has provided useful data in the past and which are essential to providing vital societal benefits. We have already acknowledged that there is a trade-off of privacy versus accuracy and that at no level of epsilon except zero is there perfect protection. Therefore, if the mathematics can't solve the issue of satisfying the documented use cases for which Census data is used, we should consider sliding along that Arc to somewhere that in combination with the mathematics does.

My final two cents. Glad to know POP is looking into it.
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

From: Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Sent: Thursday, July 9, 2020 10:30 AM
To: James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>; Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>; Kathleen M Styles (CENSUS/ADDC FED) <kathleen.m.styles@census.gov>
Cc: Deborah Stempowski (CENSUS/ADDC FED) <Deborah.M.Stempowski@census.gov>
Subject: Re: Silver Lining decades ago

Thanks James and Jane.

I believe yesterday was the first POP heard about the suggestion as it was only presented to Tori, John and me on Tuesday afternoon. I separately told Tori we should do "our homework" first before making these types of decisions and she agreed. I'm not sure she spoke to John yet.

So getting these concerns down on paper and communicating PRIOR to the next Sprint planning is critical.

Thanks,

Cynthia Davis Hollingsworth
Program Manager, 2020 Census Data Products and Dissemination
Decennial Census Management Division
U.S. Census Bureau
Office: 301.763.3655
iPhone: 202.253.6334
E-mail: cynthia.davis.hollingsworth@census.gov

From: James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>
Sent: Thursday, July 9, 2020 9:48 AM
To: Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>; Kathleen M Styles (CENSUS/ADDC FED) <kathleen.m.styles@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Cc: Deborah Stempowski (CENSUS/ADDC FED) <Deborah.M.Stempowski@census.gov>
Subject: Re: Silver Lining decades ago

I chatted with Roberto about it this morning and he said they are looking into the DEMO use cases and the effect of this tract only plan.

****
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: Jane H Ingold (CENSUS/DCMD FED) <Jane.H.Ingold@census.gov>
Sent: Thursday, July 9, 2020 8:04 AM
To: James Whitehorne (CENSUS/ADDC FED) <James.Whitehorne@census.gov>; Kathleen M Styles (CENSUS/ADDC FED) <kathleen.m.styles@census.gov>; Cynthia Davis Hollingsworth (CENSUS/DCMD FED) <cynthia.davis.hollingsworth@census.gov>
Cc: Deborah Stempowski (CENSUS/ADDC FED) <Deborah.M.Stempowski@census.gov>
Subject: Silver Lining decades ago

The decade that the school districts summary levels were added to the Redistricting PL 94-171 Summary File may be the first silver lining. We will be publishing ***total pop and the OMB race/Hispanic data at the block level and School Districts.*** Just not age and gender. I am trying to reconcile my concerns starting with Title 1 as I learned its application at the CNSAT conference. As for other uses for block data, HUD Block Grant and environmental and health, I will defer to a "process of consultation" that I urge to take place other than the current roster of CNSTAT "expert panelists" negotiating block use cases and the tradeoffs. Yesterday, the whole compressed proposal escalated to the Deputy Director in lieu of input from policy, legal, Decennial leadership, etc.

See attached use cases from POP in 2019. Look at the page SF1 block.

Jane Ingold, Special Assistant,
Decennial Census Management Division, U.S. Census Bureau
Office 301.763.4646 2H172A Cell 301.775.7515
jane.h.ingold@census.gov
census.gov Connect with us on Social Media

EXHIBIT 30

The Challenge of Invariants and the Microdata Requirement

Philip Leclerc
On behalf of & with the support of the 2020 Decennial Census Disclosure Avoidance System (DAS) development team

Data Stewardship Executive Policy Committee: Workshop in Advance of Decisions on Disclosure Avoidance
U.S. Census Bureau
August 19, 2020

Invariants are statistics the DAS can't alter

- **Invariants and microdata in the DAS:**
  - The DAS produces the MDF: *MDF = DAS(NoisyMeasurements(CEF), Invariants(CEF))*
  - Where the MDF must be microdata that satisfies: *Invariants(MDF) = Invariants(CEF)*
  - And "*microdata*" means *"a set of housing unit, household, or person records of specific types"* which may be arranged, for example (using Persons to illustrate), as a .csv file with one column for each Person variable and one row for each person
- **Currently proposed invariants:**
  - *Total Population* (by *State and State-equivalents*)
  - *Total Housing Units* (by *Census block*)
    - *Not* an invariant on number of persons in Housing Units
    - *Not* an invariant on vacant housing units
  - *Number of Group Quarters facilities by 3-digit type* (by *Census block*)
    - *Not* an invariant on number of persons in these Group Quarters
- **Caveat:** given recent re-focus on PL94-171, these may need to be re-visited on later returning to the DHCH, DHCP products

Examples of proposed invariants' implications

- If, in the CEF, a Census block has no Federal Prisons, then in this block the MDF cannot contain persons residing in Federal Prisons
- If, in the CEF, a Census block has 3 college dormitories, then in this block the MDF must contain at least 3 persons residing in college dormitories
- If, in the CEF, a Census block has only Housing Units and no Group Quarters, then in the MDF all persons in this Census block must reside in Housing Units
- Note that some quantities are constrained independent of the invariants. For example, the number of "4 year old Householders" must be 0 in the MDF; these restrictions we sourced primarily from the CEF Edit Constraints; these are called structural zeros, not invariants

Invariants weaken and complicate the privacy guarantee

- Differential privacy requires that all output statistics be subject to noise injection
- Invariants don't allow this, so the DAS does not satisfy differential privacy (but we treat invariants as the exception, rather than the rule; use of many invariants in the 2010 data made reconstruction attacks especially straightforward)
- As a result, the privacy guarantee we can make is qualitatively weaker than with full DP:
  - We cannot say: *"An attacker will learn no (very little) more about you than they could have in a world where your Census data was replaced with an arbitrary record"*
  - We can say: *"An attacker will learn no (very little) more about you than they could have in a world where your Census data was was replaced with arbitrary values, except whether you reside in a GQ (and what type), and which state you live in"*
  - We can also say a bit more about HHGQ data and state location, specifically, but this is the basic idea

Invariants & microdata can create impossible problems [A]

- Internally, the DAS doesn't operate on microdata records directly. Instead, it counts how many records there are of each possible type, and operates on these "histogram" counts

- For a structure like this, the microdata output requirement translates to requiring that, after infusing noise into the histogram counts, the resulting values must be nonnegative integers

- And the invariants translate to requiring that, for example, summing over the histogram cells that correspond to persons in Virginia must match the CEF Total Population count in Virginia

Invariants & microdata can create impossible problems [B]

- To generate microdata with accuracy that is as good as possible given a fixed privacy-loss budget, the DAS injects noise into DP measurements taken over the CEF histogram
- The DAS then finds a nonnegative, integer-valued histogram that exactly matches the invariants, and tries to make these nonnegative integer values closely match the DP measurements
- The process of "closely matching" is performed using a mathematical technique known as *mixed-integer linear programming* (MILP). With MILP, it is *very* easy to create impossibly difficult computational problems by adding invariants
- Much of the DAS team's research has focused on ensuring that current invariants do not create such a situation

Invariants & microdata can create impossible problems [C]

- Even while avoiding outright impossible problems, the DAS must decompose the ideal optimization problem it seeks to solve into simpler sub-problems
- This is necessary because privacy comes not just from which records are present, but also from those that are absent, so the DAS must consider all possible record types
- Doing this over all Census geographies at once is too much for even a very large single computer to store in memory
- To avoid this obstacle, the DAS breaks up its problem into many smaller chunks, and works on these separately (by creating a separate problem for each geographic unit in the “spine”)
- Some invariants make this decomposition difficult or impossible (e.g., Census Place total populations invariant (off-spine), but not Census Block total populations (on-spine))

Microdata creates bias, and harms accuracy

- When the DAS injects noise, the average noise value is zero. However, converting these noisy measurements – which may be negative, especially when CEF counts are small (which is common) – into microdata creates "positive bias" in small counts, and "negative bias" in large counts
- The requirement that the MDF be microdata implies that, if we calculate the on-average error in every tabulation that the DAS is responsible for producing, the worst of these on-average errors will increase (logarithmically) with the number of published tabulations
- This is not true if we just estimate DP tabulation values, not microdata: in that case, the worst on-average error depends only on the privacy-loss budget, not on the number of tabulations
- The point of this slide is that the microdata requirement is responsible for bias and additional variance in the final 2020 Census data products that are not the result of differential privacy
- We are not asking DSEP to relax the microdata requirement, just to be aware that it complicates the disclosure avoidance unnecessarily and should be thoroughly reviewed when designing new tabulation systems

EXHIBIT 31

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

To: Steve Dillingham
Director
U.S. Census Bureau

From: Allison Plyer
Census Scientific Advisory Committee (CSAC) Chair

Subject: Recommendations and Comments to the Census Bureau from the Census Scientific Advisory Committee Fall 2020 Virtual Meeting

Key
ADCOM
ADDC
ADDP
FLD
PCO
PPSI
ADRM
ADEP

September 18, 2020

The Census Scientific Advisory Committee (CSAC) thanks the 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 given 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 (ADDC) (ADCOM)

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.

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

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.

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.

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

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

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 de-duplication procedures for college students, retirees, and others require additional time.
- 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.
- Untested post data collection processing systems may fail in ways that the Census Bureau cannot foresee today.

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.[1] 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.

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.

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.

[1] https://assets.documentcloud.org/documents/7207428/LUPE-Sept-11-2020-Declaration-of-Albert-Fontenot.pdf

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

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.

Administrative Records Use (DSSD)

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

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.

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.

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[2] 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).

[2] Family Educational Rights and Privacy Act

5) 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.

6) CSAC recommends that the Census Bureau explore obtaining birth records for the first three months of the year to facilitate roster building.

7) CSAC recommends exploring whether the savings from AR use in NRFU can be applied to improve on-the-ground enumeration.

8) CSAC recommends exploring whether data on cell phones, given their ubiquity, can be used to improve counting of the hard-to-count populations.

Differential Privacy (ADRM)

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: 2020 Disclosure Avoidance System Updates. 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.

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.

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.

1) 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).

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.

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.

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.

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.

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.

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.

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 privacy-accuracy 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)

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.

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.

EXHIBIT 32

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

To: Steve Dillingham
Director
U.S. Census Bureau

From: Allison Plyer
Census Scientific Advisory Committee (CSAC) Chair

September 18, 2020

Subject: Recommendations and Comments to the Census Bureau from the Census Scientific Advisory Committee Fall 2020 Meeting, September 18, 2020

The Census Scientific Advisory Committee (CSAC) thanks the 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 given 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.

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

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.

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.

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.

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 de-duplication procedures for college students, retirees, and others require additional time.
- 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.
- Untested post data collection processing systems may fail in ways that the Census Bureau cannot foresee today.

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.[1] 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.

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.

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.

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.

[1] https://assets.documentcloud.org/documents/7207428/LUPE-Sept-11-2020-Declaration-of-Albert-Fontenot.pdf

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

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.

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

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.

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.

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[2] 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).

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.

CSAC recommends that the Census Bureau explore obtaining birth records for the first three months of the year to facilitate roster building.

[2] Family Educational Rights and Privacy Act

CSAC recommends exploring whether the savings from AR use in NRFU can be applied to improve on-the-ground enumeration.

CSAC recommends exploring whether data on cell phones, given their ubiquity, can be used to improve counting of the hard-to-count populations.

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: 2020 Disclosure Avoidance System Updates. Some aspects of the Bureau's differential privacy efforts are less transparent. 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.

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.

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).

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.

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.

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.

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.

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.

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

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.

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.

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 privacy-accuracy 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. 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)

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. 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.

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. CSAC recommends that the Bureau should delay additional releases after the December apportionment release to allow time for these recommended analyses.

Agreed prior text

Post-Enumeration Methodology

CSAC would like to thank the Census Bureau for the presentation on the Post-Enumeration Survey (PES), its history, and its basic design. CSAC appreciates the importance of the PES in evaluating the quality of 2020 Census by measuring coverage and errors, estimating overcounts and undercounts, and identifying content errors for specific questions. CSAC also agrees that the PES is an important tool to understand which methodologies worked better and worse during 2020 Census and, therefore, document valuable information for future surveys and for Census 2030.

The characteristics of PES listed above are true for any census, but the circumstances for the 2020 Census are unusually challenging. Added to the expected challenge of declining self-response rates[3], the 2020 Census has faced additional challenges from a pandemic, natural disasters, late changes to processes, and accelerated timetables. This unprecedented combination of challenges makes the importance of the PES even greater than for other decennial censuses. At the same time, the challenges affecting the 2020 Census are likely to also affect the PES.

[3] Czajka, J. L., & Beyler, A. (2016). Declining response rates in federal surveys: trends and implications (background paper). Mathematica Policy Research.

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

CSAC recommends that the Census Bureau provide maximum possible transparency on process indicators to the PES. Such transparency will increase trust in the process and may enable the Census Bureau to obtain useful feedback and help throughout the process.

CSAC would like to know a schedule for the PES (for conducting the survey, processing data, and releasing it) to the best degree it is known at the time of the Census Bureau's response. CSAC would like to know if the Census Bureau is considering any methodological adjustments to the PES given the unprecedented context. For example, any changes to how respondents are contacted or any consideration of re-weighting the sample observations?

CSAC also recommends that the Census Bureau consider how to separate lessons learned from the PES from those related to temporary challenges (e.g., the pandemic), versus ongoing challenges (e.g., lower self-response rates).

In view of the challenges faced by the PES, which is just beginning field data collection, CSAC requests that the Census Bureau provide a detailed update at its spring 2021 meeting so that CSAC members can review the Bureau's progress and have an opportunity to offer suggestions to address outstanding methodological issues, including development of a suitable correction for correlation bias among children.

Pulse Surveys

CSAC commends the Bureau for the incredible initiative and nimbleness in fielding the Pulse surveys. Large scale (including regional) disasters create an enormous break in the status quo followed by months and often years of flux. By their very nature, disasters create confusion and rumors start to circulate post-disaster. After a disaster there is escalated demand for timely data to address rumors, assess the current status, and re-assess over time—all to inform and catalyze critical community, planning, and investment decisions that support recovery. Indeed, post-disaster data can actually catalyze investments by reducing uncertainty, and as such is a very important contributor to recovery. But just as demand for data escalates after a disaster, leaders complain of a vacuum of data after a disaster. This is because most data is insufficiently timely to be relevant after a disaster.

The Bureau's work to address pressing questions in a timely manner soon after the COVID crisis began is an extremely important example of the kind of data collection that is essential following each disaster. Updating this data frequently is critical as well. As conditions change, the questions may change, and the frequency with which data is collected can decrease over time. The Bureau is learning these lessons about the dynamics of post-disaster data demand through their own direct experience.

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

CSAC commends the Bureau for considering institutionalizing rapid response surveys for future national or regional emergencies. Such efforts can go a long way to forwarding disaster recovery. While nearly all large-scale disasters are similar in generating demand for data, disasters often differ in the types of data needed.

The Data Center of Southeast Louisiana uses a version of Design Thinking that the Bureau could consider implementing. It entails scanning local media for descriptions of lived experiences (e.g. small businesses closing, people unable to return to their homes, workers becoming disconnected from employers) combined with demographic and economic expertise to develop hypotheses about whether such experiences are likely to be common or widespread. The Data Center prioritizes gathering data on those experiences or issues that specifically can inform recovery, planning and investment decisions.

Gathering local knowledge about decisions that are being made with no data or bad data, would be another way to identify and prioritize data collection post-disaster. The response period is short-lived and the Bureau's data collection efforts are likely to be more impactful if focused on informing recovery efforts rather than immediate response. The Bureau could reach out to local planning departments of disaster-affected municipalities to identify the decisions they are struggling to make because of lack of data.

To create a question bank in advance, the Bureau could assemble a number of leaders who have hard-earned, on-the-ground expertise in disaster recovery following common disasters such as wild fires and hurricanes to identify some of their most common data needs. The National Low Income Housing Coalition currently convenes a group of such experts on a regular basis. Resilient Cities Catalyst is a Rockefeller Foundation initiated organization that supports recovery and resilience across many cities and could assemble a group of experts to identify high impact data needs. School districts, whose operations were particularly hard hit by COVID, could also be consulted. Many cities have Chief Resilience Officers who could provide good inputs. The Bureau could also get input from long-term resilience committees and longer-term recovery committees that have now be established by many states.

CSAC recommends the Bureau consider a TOP (The Opportunity Project) Sprint that engages diverse stakeholders recovering from disasters. Because individuals from the most vulnerable communities are almost always the most deeply impacted by disasters, TOP could specifically engage representatives from the such communities -- particularly those with Access and Functional Needs, lower income communities, households with language barriers, renters, etc.

Because it is based on the MAF, the Household Pulse Survey can be linked to other Census data like the American Community Survey. When resources permit, CSAC recommends creating a (restricted) version of the HPS matched with other data at a fine geographic level (e.g. Census tract) for each respondent. This would allow researchers to understand which characteristics of communities helped predict relative success or failure in confronting the pandemic.

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

The pandemic (and other shocks) are being experienced differently in and out of metropolitan areas. Reliance on public transportation and the quality of medical facilities, for example, differ in metropolitan and non-metropolitan areas. At present, the Household Pulse Survey is representative for states and for the 15 largest MSAs. If resources permit, CSAC recommends making future Pulse Surveys representative for MSA status by state (or at least by Census region).

The weekly frequency of the Pulse data is touted as an advantage. With the two phases nearing completion, CSAC recommends formal tests for the incremental value of having the data at a weekly rather than monthly frequency. Indicators from the Pulse Surveys presented in the one-way briefing in August did not, in general, reveal sharp week-to-week differences.

CSAC further recommends that the Pulse Surveys include some identical questions to other existing administrative data sets and surveys being conducted on the national, state, and local levels, as a means of cross-validating the Pulse Surveys. At the national or state level, such administrative data sets and economic surveys might include Initial UI Claims (weekly), Current Population Survey (monthly), Current Employment Statistics (monthly), and ISM Reports on Business (monthly). The need for cross-validation is particularly important given the understandably low response rates on the Pulse Surveys. Once validated, the Pulse Surveys can be used to augment the key elements of those existing surveys on a timelier basis during times of national or local emergency. Further, the wording of questions in the Pulse Survey could provide a template for state and local jurisdictions to implement their own surveys as local conditions warrant.

Both Pulse surveys emphasize questions whose answers would be hard to obtain without household interviews. This is the comparative advantage of the Pulse surveys relative to other datasets that have been used to study COVID's effects. Since so much of any given household's experience depends on how the whole community responds to COVID-19, CSAC recommends including questions in the Household Pulse survey on the household's perception of the community's adherence to wearing masks, social distancing, quarantining, etc. Questions on internet access could also be added, beyond its role in primary and secondary education, given its importance during the lockdown periods. Additionally, since so much of the federal government's assistance to small businesses was in the form of potentially forgivable loans administered through banks, the health of the local banking sector is important to understand. CSAC recommends that future instances of the Small Business Pulse Survey include questions about the performance of the banking sector in administering these programs.

The Pulse Surveys reflect the Census Bureau's comparative advantage in data collection – direct surveys of representative samples. There have been many other data collection efforts, relying more often on passively collected data that are not necessarily representative of the total

population. Some examples include transaction-level data from financial intermediaries, social network data aggregated by locality, and smartphone location data. CSAC recommends that when time and resources permit, the Census Bureau should compare the conclusions drawn with each type of data and integrate them into a comprehensive report on the pandemic.

Lastly, CSAC commends the Bureau for offering the Pulse survey in Spanish. However, the instructions for Spanish speakers are in English, which may be an obstacle for Spanish language speakers. CSAC recommends that the Bureau make the instructions as well as the Pulse survey itself available in Spanish.

Construction Modernization

CSAC would like to thank Stephanie Studds and her team for such impressive work. The use of innovative techniques, such as the change detection from satellite images, allows the team to modernize the collection and analysis of construction data.

The goal of the Construction Modernization project is to reengineer the measurement approach to the traditional construction surveys by utilizing alternative data sources, developing modeling techniques, and evaluating the use of satellite technology.

The Economic Indicators Division of the Census Bureau has been tasked with the modernization effort and has made significant progress modernizing the workflow. The Working Group's focus is to work with the Census Bureau to provide input to assist with the following:

- Developing methodologies to maximize data consistency, accuracy and geographic coverage and granularity
- Reducing the need for field collection and/or burden to respondents
- Implementing a methodology for real-time data ingestion and updates
- Implementing a methodology that remains cost neutral across the construction programs
- Defining key milestone markers, such as what indicators show completion of construction
- External communication related to the impact of the program
- Communication of findings and reports

The Construction Modernization Working group currently consist of four CSAC members, all of which will complete their term on the CSAC in Spring of 2021. There is a need for additional CSAC members to participate in the working group to provide continuity and a means to transition the effort.

These activities make it evident that partnerships with the private sector, particular industries of interest such as insurance and housing, could enhance the expected products. Other Census

data products could also benefit from these activities. We encourage internal communication and coordination among the Census. For example, the tempo of the urban/rural classification of the Census blocks could be enhanced by the utilization of these data.

CSAC recommends that the Census Bureau continues the Construction Modernization – Re-engineering Initiative and includes a follow-up presentation of results at the next CSAC meeting.

2020 Census CVAP Special Tabulation

CSAC commends the effort that went into securing access to administrative data from various sources, matching those records, and formulating and testing various methodologies for imputing citizenship for those records that either did not link or otherwise could not be assigned citizenship status from administrative records.

CSAC recommends that the Census Bureau provide a summary breakdown of its citizenship estimates into four subgroups:

1. BR citizen
2. BR non-citizen
3. Imputed citizen
4. Imputed non-citizen

and present these results for each of the four modeling methods and, if available, for both the 2010 CEF and the 2018 ACS.

Prior to making more substantial recommendations or responding to questions posed by the Census Bureau, it would be helpful for CSAC to see a draft of the technical document planned for October release, so that we can more carefully review and understand the modelling process and the context behind the results shown in the presentation.

CSAC recommends the Census Bureau provide justification for why CVAP data need to be produced down to the block level, especially given the viability of producing accurate estimates at such a small unit and with differential privacy applied. Currently, estimates of deviation between DP and SF1 2010 data at the block level show MAPE (mean absolute percent error) in the range of 50-78% for total populations. This deviation would be increased substantially with 12 race/ethnic breakdowns and restricting to the citizen voting age population.

Questions:

- We have seen higher PIK rates for the ACS using the most advanced PVS methods, how comfortable is the Census Bureau with the 91% match rate?
- If citizenship is missing or outdated in the SSA data, to what extent were they able to fill this with passport and naturalization records? Are the naturalization records complete?

Recommendations and Comments to the Census Bureau
from the Census Scientific Advisory Committee
Fall 2020 Meeting

Public Comments

CSAC appreciates the Census Bureau enabling public engagement and recommends that the Census Bureau respond in writing to the written public comments from Joseph Battistelli submitted via the chat feature and the Deborah Stein submitted via a letter.

EXHIBIT 33

DRAFT//DELIBERATIVE//PRE-DECISIONAL//CONFIDENTIAL//CUI

Draft of Memo about Concerns with Intentionally Distorting the Population Tabulations in the P.L. 94-171 Redistricting File

The U.S. Constitution mandates that an "actual Enumeration" (Article I, section 2) "counting the whole number of persons in each State, excluding Indians not taxed" (Amendment 14, section 2) be conducted "within every subsequent Term of ten Years, in such manner as they [Congress] shall direct" (Article 1, section 2). Congress, in turn, has directed that the above "decennial census of population" be utilized to produce both a "tabulation of total population by States…as required for apportionment" as well as within "geographic areas for which specific tabulations of population are desired" by the "officers or public bodies having initial responsibility for the legislative apportionment or districting of each State" with the Secretary retaining "final authority for determining the geographic format of such plan" (Title 13, §141) The need to conduct an "actual Enumeration…counting the whole number of people in each State, excluding Indians not taxed" and to produce those counts at both the State-level for apportionment and at more specific levels of geography within States based on plans coordinated between the Secretary and the State governments for potential use in redistricting are statutorily required, and underpin the delicate political arrangements that have been constructed over centuries to equitably distribute political representation and, based on other statutes, federal funding. In addition, §141 of Title 13 also specifically states 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."

With a novel legal interpretation, some in the U.S. Census Bureau have publicly and repeatedly indicated that in order to now comply with §9 of Title 13 to not "make any publication whereby the data furnished by any particular…individual under this title can be identified," they must, for the first time in U.S. history, now distort the basic population tabulations required in §141 via an algorithm so that States do not receive the actual population tabulations that were counted in the "decennial census of population" for the geographic areas they have coordinated with the Secretary. The argument is basically that in order to comply with aspects of §9 of Title 13, they must no longer comply with aspects of §141 of Title 13.

This is a false choice. The population total for a census block is not itself "data furnished by a particular…individual." Protecting the "data furnished by a particular…individual" does not mandate or authorize the intentional distortion of population totals that are required by §141 of Title 13. In pursuit of its self-described mission to "ensure that the Census Bureau protects Title 13 respondent confidentiality," the Disclosure Review Board (DRB) at the U.S. Census Bureau has no authority to mandate or authorize that basic population tabulations required by §141 of Title 13 be intentionally distorted via an algorithm.

Not only would the intentional distortion of population counts via an algorithm be a violation of §141 of Title 13, it would also "not solve the global problem of personal data disclosure" (see the CIC/FSCPE/SDC letter), cause harm to various local governments whose population counts are randomly and arbitrarily decreased (see the State of Maine letter), create

systemic "positive biases, particularly in small domains where rounding up occurs to avoid negative values" (see the CSAC memo), create internal litigation issues for States (see the NCSL letter), and cause "potential unintended consequences… on the allocation of important funding and other activities that rely on census data" (see letter from 33 Congressional members).

Indeed, the Census Bureau has received a high volume of concerns about the proposed intentional distortion of population counts from a variety of parties who are reliant on the accurate production of population counts as determined by the decennial census, as required by §141 of Title 13. The following is a sample of concerns received by the Bureau:

In a letter dated November 27th, 2019, the Census Information Centers (CIC), Federal-State Cooperative for Population Estimates (FSCPE), and State Data Centers (SDC) stated the following:

- "We have concerns that this implementation has been driven by data scientists with limited consideration for users' needs."
- "We are particularly concerned that insufficient analysis has been conducted regarding how DAS will affect the Census data used for informing policy and allocating public and private funds."
- "We have concerns that the proposed implementation violates the Census Bureau's obligation, under 13 USC section 141, to provide a Redistricting Data File with accurate population counts."
- "There is considerable concern about Differential Privacy DAS, its origins as a policy, and its implementation."
- "It is important that our network members, data users, and the stakeholders we serve understand why the Bureau took the action proactively to be the global leader in disclosure avoidance without, as far as we can know, any major challenge that the Bureau was not upholding its 13 USC mandate. It is also important to acknowledge that the Bureau's initiative will not solve the global problem of personal data disclosure."

In a letter dated January 17th, 2020, the Maricopa Association of Governments wrote:

- "We have reviewed the Census Bureau's differential privacy proposal and have concerns that the proposed differential privacy methodology would have negative consequences in allocation of federal and state revenue, redistricting of legislative, congressional, and city council districts, and would cause significant issues related to planning for transportation, other types of infrastructure, and, human services."

In a letter dated February 3rd, 2020, the City of Alexandria, Virginia wrote:

- "We are concerned that the proposed differential privacy methodology would limit our understanding of the city's population, and inhibit our ability to serve our residents equitably."

- "If Census data do not reflect reality, the system could unintentionally be designed to overserve some communities and underserve others."
- "…accurate block-level data are critically important to the City's understanding of the current population and ability to anticipate future population growth."

In a letter dated February 13th, 2020, the Utah State Legislature wrote:

- "… we fear that differential privacy will require the states to legally defend whether differential privacy protected census data will satisfy the states' constitutional obligation to meet population and equality requirements."
- "…the integrity of the data used to redistrict the state into congressional and legislative districts…will be threatened."
- "It is therefore our recommendation that the bureau increase its efforts to hold the census block population data invariant."

In a letter dated February 20th, 2020, the Department of the Administrative & Financial Service of the State of Maine wrote:

- "The U.S. Census Bureau has long been the standard-bearer in terms of providing high quality, reliable data to the public. This proposed policy change would throw into doubt any redistricting, funding decisions, or analysis done using census data."
- "DP has the potential to exclude rural and resource-strained communities from equitable access to high-quality, reliable data, and that our narratives will be systematically misinformed as a result."
- "This will have myriad financial and economic repercussions for the 'winners' and 'losers' that municipalities will randomly become."

In a letter dated May 21st, 2020, the National States Geographic Information Council (NSGIC) wrote:

- "…new practices will negatively impact state programs and the ability to carry out statutory responsibilities, in effect harming the citizens the DAS aims to protect."

In a letter dated March 26th, 2020, the National Conference of State Legislatures (NCSL) wrote:

- "States are required to comply with the U.S. Constitution's 'one-person, one-vote' principle and with protections provided by the Voting Rights Act of 1965 (as amended). If block-level census data is released in a form that is known to not represent the actual number of people enumerated at the block level, states may find themselves litigating based on the quality and accuracy of federal census data before plans are drawn and even afterwards."

In a letter dated June 25th, 2020, the National Congress of American Indians (NCAI) wrote:

- "We have clearly stated in multiple meetings with the U.S. Census Bureau since last year that the 2020 Census data must be accurate for the following priority use cases: 1) reapportionment and representation; 2) federal funding formulas and decision-making; 3) local tribal governance; and 4) AI/AN research and surveillance data. U.S. Census Bureau staff informed NCAI and tribal leaders in the Census Roundtable Discussion that a new geographic spine strategy would be tested to address the priority use cases for political and legal entities and would place AI/AN data on the geographic spine, make AI/AN data within a state invariant, and would give AI/AN geographies their own direct allocation of the privacy loss budget. While we were interested to see how this proposed plan would fare in Sprint II, we were recently informed that the new geographic spine was not tested and instead was dismissed by U.S. Census Bureau officials as "too hard" to implement. While our team was repeatedly assured by U.S. Census Bureau staff in numerous calls, meetings, and virtual workshops since January that our concerns were being addressed, we recently learned the information that was provided to us was in fact, not true. We are losing confidence in your efforts to make adjustments to the DAS and are concerned that our priority use cases are not being addressed. Even your own staff admitted after the December 2019 National Academy of Sciences, Engineering, and Medicine workshop that the results of applying the DAS to the 2010 demonstration product were "unacceptable." It is now six months later, and the results of Sprint II are even more inaccurate.

In a letter dated August 24th, 2020, the Co-Chairs of the Congressional Native American Caucus wrote:

- "It is critical that the 2020 census includes accurate data for tribal communities for the purposes of representation, reapportionment, federal funding formulas, accurate research, and tribal government planning and service delivery…"

In a memo dated September 18th, 2020, the Census Scientific Advisory Committee (CSAC) wrote:

- "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."
- "…many implications of this decision for privacy, accuracy, and fitness-for-use are currently unknown."
- "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…CSAC recommends that the Bureau should delay additional releases after the December apportionment release to allow time for these recommended analyses."

In a letter dated September 21st, 2020, 33 members of the House of Representatives wrote:

- "We write to express concern with the U.S. Census Bureau's proposed "differential privacy" approach to maintain the confidentiality of data collected in the 2020 decennial census… The "differential privacy" method proposed by the Census Bureau involves injecting statistical "noise" into data at the sub-state level, such as at the region, district, town, and census block levels, prior to its release, altering the count and characteristics of individuals and households reported at these levels… As Members of Congress, we are concerned about the potential unintended consequences that a differential privacy method could have on the allocation of important funding and other activities that rely on census data. The United States has a rich and diverse economy and workforce, and the needs vary widely across our states and congressional districts, which is why it is essential that sub-state level census data be accurate and reliable."

It is therefore recommended that population totals in the 2020 census be held invariant for all levels of geography agreed upon between the Secretary of Commerce and the "officers or public bodies having initial responsibility for the legislative apportionment or districting of each State" and that the Census Bureau's Data Stewardship Executive Policy Committee (DSEP) and Disclosure Review Board (DRB) be informed of their lack of authority to mandate, contrary to law, that population totals for either apportionment or redistricting be intentionally distorted via an algorithm. Instead, Census Bureau should consider other options to maintain compliance with §9 of Title 13.

EXHIBIT 34

FY 2021 DOC Senate QFRs

CENSUS

QUESTION SUBMITTED BY SENATOR LISA MURKOWSKI

CENSUS

Background

The Census Bureau will be implementing a new privacy system, known as 'differential privacy,' for the first time starting with the 2020 U.S. Decennial Census. We know that the method is meant to balance data accuracy with data privacy. However, I am worried about the impact to small population groups such as American Indians and Alaska Natives. A recent analysis by Randall Akee, (presented at the National Academy of Sciences https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518) of certain Alaska Native villages found that applying differential privacy could result in a nearly 30% undercount of American Indian/Alaska Native people in 2020 compared to Bureau's data sets from the 2010 U.S. Decennial Census. This could thus impact the US government's trust responsibility, which is grounded in the U.S. constitution and federal-tribal treaties. Many agencies meet this responsibility and often rely on the Bureau's data for their formula allocations. An undercount will potentially reduce a tribe's funding. I believe that the Bureau is aware of this problem but I want to ask you about progress in the development of privacy systems that meet user needs as well as protecting confidentiality.

Question A: Does the Bureau have enough time to develop and implement another privacy system, given that the 2020 Census has started? How much undercount of Alaska Native villages is acceptable, given that the data are essential for helping the government meet its federal obligations to tribes?

The Census Bureau is confident that we will successfully deploy the 2020 Disclosure Avoidance System (DAS) on time to produce high quality data products while protecting the privacy of our respondents as required by law. Faced with significantly greater privacy threats this decade than the Census Bureau has previously had to protect against compels us to ensure that this modernization of our privacy protections is a success. Abandoning the DAS and retreating to the methods applied in prior decades, in a manner that would meet our legal obligations to protect privacy, would necessitate suppressing large amounts of census data from our publications and introducing significantly more noise and error through record swapping. Such an approach would render census data almost entirely unusable.

The 2010 Demonstration Data Products, which were released in October 2019, were the first production run of the DAS' TopDown Algorithm (TDA) at the quasi-full scale of the decennial census. This production run demonstrated that the algorithm can effectively protect privacy in the billions of tabulations necessary to produce the 2020 Census data products. As our data users have noted, however, in feedback to the Census Bureau, at the December 2019 National

Academies' workshop, and in the media, the 2010 Demonstration Data Products did contain a number of worrisome inaccuracies and distortions that needed to be addressed. The version of the algorithm that produced these demonstration products, however, was from eight months ago, and the DAS team has been diligently identifying and implementing solutions to address these shortcomings.

The Census Bureau works diligently to minimize the occurrence of undercounts for any population group, including the AIAN population. To that end, the Census Bureau does not consider any undercount of Alaska Native villages resulting from the application of privacy protections to be acceptable.

Improvements to the DAS will continue over the coming months, and we are working closely with a number of expert groups to assess and report out on these improvements. The Census Bureau is committed to producing high quality data, while protecting the privacy of our respondents. The 2020 Census Data Products will reflect that commitment.

Question B: The Bureau's National Advisory Committee on Race, Ethnic, and Other Populations recommended that the agency's Data Stewardship Committee use one hundred percent counts for the AIAN population for purposes of federal allocation formulas. Does the Bureau plan to provide this data for agency formulas?

The Census Bureau recognizes the special trust relationship that the United States has with federally recognized American Indian and Alaska Native (AIAN) tribes, and we understand the importance of providing accurate population counts for AIAN communities and geographies. The vital nature of this obligation is especially clear in the context of federal allocation formulas, on which many of these communities critically rely. Given the widespread use of Census data for these funding allocations, and the transparency necessary to ensure that these funds are equitably distributed, the confidentiality restrictions of Title 13, Section 9 preclude producing two sets of data—one that is publicly released, and one that is provided confidentially to funding agencies, which would be normal in this situation. Consequently, the Census Bureau is instead focusing on improvements to the DAS to maximize the accuracy of the official 2020 Census AIAN population counts, while protecting the privacy of the individuals within these communities in a single data set. In consultation with the National Congress of American Indians, the Alaska Federation of Natives, tribal leaders, and expert data users representing AIAN communities, we have developed a set of accuracy measures against which we can assess and report on the success of these improvements over the coming months.

QUESTION SUBMITTED BY SENATOR MARCO RUBIO

Census

Background

It is encouraging to hear your emphasis on ensuring that the Census is performed in a complete and precise manner. As the Senator for Florida and a member of the Special Committee on Aging, I am particularly concerned about scammers posing as Census workers to obtain sensitive information, such as credit card and social security numbers, from elderly victims.

Question A: How can Congress best assist the Department and the Census Bureau to help educate the public and combat these bad actors?

Answer A:

The Census Bureau shares your concern. Knowing that potential bad actors exist, we devoted a significant portion of our 2020 census website to advising the public on how to avoid frauds and scams[1] in preparation for our decennial count. On the site, we address the topics of avoiding scams online, the legitimate communications we send to respondents, how to verify the identity of someone who comes to your home purporting to collect a response to the census, and reporting suspected fraud.

We ask that you make your constituents aware that during the 2020 Census, the Census Bureau will never ask for information such as your Social Security number, bank account or credit card numbers, anything on behalf of a political party, or money or donations. Furthermore, they should know that all valid Census Bureau websites will always have ".gov" at the end and that the Census Bureau will not send unsolicited emails or text messages to request their participation in the 2020 Census. If they are not sure if the communication they received is legitimate or they suspect any other kind of fraud, they may also call us at 844-330-2020 to speak with one of our representatives.

In an effort to combat the spread of misinformation (incorrect information spread unintentionally) and disinformation (incorrect information spread intentionally) about the 2020 Census, the Census Bureau last year established a Trust & Safety Team. The team is committed to ensuring that the information your constituents receive about the census is factual and accurate. To this end, it monitors all available channels and open platforms for misinformation and disinformation about the census, allowing us to respond quickly to fight potential threats to achieving an accurate count in traditional media, social media and other stakeholder communications. Along similar lines, we have also launched a dedicated page on our website, Fighting 2020 Census Rumors[2], and encouraged partners and stakeholders to report anything that looked suspicious to an email account set up for this purpose (<rumors@census.gov>).

[1] See https://2020census.gov/en/avoiding-fraud.html
[2] See https://2020census.gov/en/news-events/rumors.html

QUESTION SUBMITTED BY SENATOR JACK REED

CENSUS

Background

As we discussed during the hearing, I remain very concerned with the status of the Census Bureau's targeted outreach campaign for Providence County, the only location of the End-to-End Census Test. As such, please provide an update on the Census Bureau's actions and plans to implement the targeted outreach campaign, including an answer to the following specific questions.

Question A: How much money has the Census Bureau allocated to the targeted outreach campaign as of March 12?

Answer A:

We are unable to provide media costs broken out by state, county, congressional district, or any geography below the total U.S. This information is proprietary to our contractor. Additionally, since every state, county, and city benefits from a national ad buy, there is no way to quantify the cost associated with any specific state, county, or city.

The 2020 Paid Media Campaign is $323.5M (as of April 27, 2020) and the planned budget is broken out by audience.

Audience	2020 Plan
Diverse Mass (Traditional)	$86.0MM
Diverse Mass (Digital)	$67.8MM
Hispanic	$62.3MM
Black/AA	$46.9MM
Asian	$23.9MM
AIAN	$7.5MM
Puerto Rico	$2.8MM
NHPI	$2.3MM
Emerging and Legacy	$2.8MM
New Languages	$2.7MM
Breakthrough Initiatives	$8.5MM
Contingency	$10.0MM

Question B: How much money does the Census Bureau intend to allocate to the targeted outreach campaign in total?

Answer B:

See Answer A above.

Question C: Please list every newspaper, radio station, television station, and digital platform that the Census Bureau has or will run advertisements on as part of the targeted outreach campaign.

Answer C:

Print ads specific for Providence County ran in various statewide and local newspapers and journals, including the Providence Journal and the Valley Breeze. Those began in late February and continued into late March.

Digital ads ran on various web and social media platforms, including the Providence American, Facebook, and Instagram. Digital ads targeted to Providence County started on February 19 and continued until approximately March 24. We estimate we delivered approximately 3.5 million impressions in digital ads to Providence County's residents.

Question D: Similarly, please list what languages – other than English – the outreach campaign will be implemented in and on which platform and which days those advertisements have appeared or will appear.

Answer D:

Broadcast ads on Rhode Island radio stations began on March 2 and continued until about March 29. The radio ads included English, Spanish, and Portuguese.

EXHIBIT 35

UNITED STATES DEPARTMENT OF COMMERCE
U.S. Census Bureau
Washington, DC 20233-0001

Mr. Jeff Hardcastle
State Demographer
4600 Kietzke Lane
Building L, Suite 235
Reno, NV 89502

Dear Steering Committee Members:

Thank you for your letter following up on our February 26 responses to your questions about the U.S. Census Bureau's adoption of differential privacy to protect the confidentiality of respondent data for the 2020 Census.

As we stated in our prior letter, the Census Bureau places great value in the partnership and support provided by your networks, and we appreciate your collective commitment to helping the Census Bureau meet its dual mission of producing high quality statistics about the nation, while safeguarding the privacy of our respondents and the confidentiality of their data.

In your letter, you raised six additional questions about the Census Bureau's adoption of differential privacy and the implementation of the Disclosure Avoidance System, and make three recommendations for the Census Bureau's consideration. Enclosed you will find our responses to your questions and recommendations.

Thank you for your continued commitment to a successful 2020 Census.

Sincerely,

JOHN ABOWD Digitally signed by JOHN ABOWD
Date: 2020.06.24 14:47:58 -04'00'

John M. Abowd, Ph.D
Associate Director and Chief Scientist
Research and Methodology

Enclosure

FSCPE Question #1 – "We have heard repeatedly that the decennial census has error in it for several reasons. John Abowd has emphasized this and is hiring staff to examine and quantify the sources and their contribution to that error. The underlying principle for differential privacy is that the census is accurate enough that an accurate identification of an individual can be made. Given the current delays in 2020 Operations with the current crisis and issues with address listing, etc. how can moving forward with differential privacy be justified?"

For decades, the Census Bureau has been diligent at assessing and reporting on sources of error in census counts. These sources of error include operational error, coverage error, and measurement error. Sources of error are routinely reported as part of the Census Coverage Measurement Program. In spite of these errors and the additional error introduced to protect privacy through data swapping, our internal evaluations have determined that the 2010 Census data were still accurate enough to enable confirmed re-identifications for 52 million individuals using only a portion of the published data. Our internal assessment was later confirmed by an independent, external group of scientists and data experts convened by the JASON advisory group. The JASON report on the Census Bureau's decision to adopt differential privacy for the 2020 Census states, *"In the view of JASON, Census has convincingly demonstrated the existence of a vulnerability that census respondents can be re-identified through the process of reconstructing microdata from the decennial census tabular data and linking that data to databases containing similar information that can identify the respondent."* The report goes on to state that "*in view of the demonstrated vulnerability, it is clear that the usual approaches to disclosure avoidance such as swapping, top and bottom coding, etc. are inadequate."* With the JASON's findings confirming our own internal assessments, the Census Bureau stands by our decision that the only way to meet our statutory obligations under Title 13 to protect respondent privacy is to modernize our disclosure avoidance methods through the application of differential privacy for the 2020 Census.

FSCPE Question #2 – "Does the proposed Disclosure Avoidance policy for the 2020 Census represent a new interpretation of Title 13? If so, why now?"

The Census Bureau's decision to adopt differential privacy for the 2020 Census does not reflect a change in our interpretation of Title 13. Rather, it reflects growing empirical evidence, confirmed by outside experts and our own internal researchers that the privacy risks associated with publishing large amounts of highly granular tabulations have increased substantially over the last decade.

FSCPE Question #3 – "Similarly, some DP literature talks about it as a response to a potential not actual threat. Has there been research that assesses the threat level and types of risk to the general public that DP is meant to prevent other than examples like Netflix?"

When the Census Bureau published the 2010 Census Data Products, the disclosure avoidance methods employed for their release were sufficient to protect respondent privacy at that moment in time. Within a few years, optimization algorithms had improved sufficiently to significantly increase the risk of re-identification. Recognizing that the Census Bureau cannot

rely on assumptions that the residual risk of re-identification from the application of disclosure avoidance methods will always remain constant, the Census Bureau determined that we needed to use techniques that do not rely on assumptions about what technology a would-be adversary would be able to leverage against us. Differential privacy establishes a future-proof upper bound on the leakage of private information in published data. While this privacy guarantee does represent a worst-case scenario, and thus does not necessarily reflect the actual risk of re-identification at any given moment in time, our experience has demonstrated that there is no way to predict how that actual risk will increase over time. Our conclusions on both the risk and the need for formally private solutions have been validated by external researchers, including in the JASON report referenced above.

FSCPE Question #4 – "Why, at such a late stage, is the DSEPC convinced they can create useable data with minimal input from stakeholders (given the timeframe) when their efforts to date have not provided quality data?"

The Census Bureau is engaging extensively with our stakeholders to ensure that the 2020 Census Data Products will be fit-for-use for the priority uses of census data, consistent with our obligations under Title 13. In addition to our ongoing stakeholder engagement, we are working with expert working groups organized by the Committee on National Statistics and by our National Advisory Committee and the Census Scientific Advisory Committee.

FSCPE Question #5 – "Can we be assured that the published data, including the second group of products to be released, will be internally consistent, for example that household population in the population- based tables is consistent the household population in the housing-based tables?"

All Group I (PL94-171 redistricting file, Demographic Profiles, and Demographic and Housing Characteristics file) person-level tables (P-tables) will be internally and hierarchically consistent, as will all the Group I household-level tables (H-tables). There are currently no plans to ensure consistency between the Group I P- and H- tables. Consistency between the Group I and Group II (detailed race, detailed ethnicity, tribal data, and person-household joins) data products will be established through constraints that the Group II P-tables must be less than or equal to their Group I P1 equivalent, and the Group II H-tables must be less than or equal to their Group I H1 equivalent.

FSCPE Question #6 "We appreciate the recent release of the metrics. However, it is still unclear the path forward for engaging stakeholder in a dialogue. It is still unclear what the DAS implementation plan is and so what our role is, what other groups are part of the outreach effort, and what are the deliverables and due dates. We need this information to fully inform our elected officials and impacted agencies."

The operational schedule under which the DAS team is currently working has the Census Bureau's Data Stewardship Executive Policymaking Committee (DSEP) making final decisions in September 2020 about the overall algorithm design and the final list of which data elements

will be held invariant. DSEP will then set the final privacy-loss budget and its allocation across data products, tables, and geographic levels in March 2021. Stakeholder feedback to inform those decisions will be invaluable to DSEP's decision-making. To that end, we are currently actively engaging with our stakeholders through various channels, including working groups organized by our advisory committees, by the Committee on National Statistics, and by FSCPE, as well as ongoing engagement with American Indian and Alaska Native tribal leaders and data users, among others. In particular, we are asking these groups to provide feedback through this summer about how we are assessing accuracy and identifying the priority use cases of census data to help inform the September 2020 DSEP decisions. Similarly, feedback that we receive from these groups on minimum acceptable thresholds for accuracy to support various priority use cases, and suggestions for communication materials and supporting guidance on fitness-for-use will help inform the March 2021 DSEP decisions.

FSCPE Recommendation #1 – Provide a clear, consistent and timely communications plan for keeping the full range of stakeholders informed about this. In the letter, Dr. Abowd provides a general overview of outreach efforts. However, we continue to find that the squeakiest wheels are getting the Bureau's attention. This appears to be the case for privacy advocates as well as those of us concerned about data accuracy.

Improvements to the DAS are occurring on a continuing basis, and we recently have expanded our traditional communications channels and enhanced our ability to get timely information out to the diverse data user community by establishing an email newsletter. Data users can subscribe to get prompt notifications regarding important DAS developments. We also post all of our updates on our DAS Updates webpage. Data users that want to provide feedback can do so by emailing our DAS and Data Products teams at 2020DAS@census.gov.

FSCPE Recommendation #2 – As soon as possible, we need information on how the Bureau will provide information on the noise infused products. We need information on whether not DP can be treated as a random error and the ranges of that error. We need to be able to inform our state and local administrative and legislative bodies on changes that will materially impact their operations.

We thank you for this recommendation. The Census Bureau is committed to providing fitness-for-use guidance for our 2020 data products. What form that guidance will take is still under consideration. Similarly, any guidance on error ranges or similar metrics will necessarily need to wait until finalization of the DAS algorithm design and its parameters (i.e., the final privacy-loss budget and its allocation). In the interim, we welcome any feedback that our data user community would like to provide regarding what form this guidance should take and what it should contain. Please send this feedback to Michael.B.Hawes@census.gov and 2020DAS@census.gov.

FSCPE Recommendation #3 – Most importantly, we believe that any further iterations of demonstration data run through a revised DAS are necessary for review. We need to be able to rerun our earlier evaluations against any revised data. Researchers from the CNSTAT

meeting also reached this conclusion. The Demonstration data product has proven to be both useful and necessary to ensuring both the quality and utility of the final release product. As the implementation of the DAS system will impact the data for everyone in the country the need for through, rigorous, independent review of the data is obvious. This review must be a full review of the data not simply based on metrics.

The accuracy metrics that we have developed are intended to allow our data users to assess our ongoing improvements to the DAS algorithm and their impact on fitness-for-use in a variety of ways. That said, we recognize that for some important uses of census data there is no substitute for actually examining the underlying data. You recommend that the Census Bureau release additional demonstration data products to support in depth analysis of the data's fitness-for-use. Unfortunately, the tabulation, documentation, and quality control processes that the Census Bureau employs for public releases of data products are enormously time and labor intensive. With the 2020 Census now underway, we are unable to support additional releases at the present time. That said, in order to support these detailed assessments without overburdening our tabulation and data products teams, the Census Bureau is committing to release the differentially private, but untabulated Privacy-Protected Microdata File (PPMF) produced by each successive iteration of the DAS algorithm for which we release new Detailed Summary Metrics. While these PPMFs will not be in the standard table structures associated with the PL94-171 or DHC data products, it would be an easy matter for some of our public data users to tabulate them accordingly. The PPMFs provided will be exactly the same data used to prepare the Detailed Summary Metrics. They are also exactly the same data the Census Bureau would have tabulated into new demonstration data products. We trust that this solution will meet your needs.

EXHIBIT 36

Expert Report of Michael Barber in Reply to Amicus Brief of Data Privacy Experts

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

26 April 2021

1 Introduction and Qualifications

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

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

I have worked as an expert witness in a number of cases in which I have been asked to perform and evaluate various statistical methods. Cases in which I have testified at trial or by deposition are listed in my CV, which is attached to the end of my initial report, dated March 9, 2021.

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

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

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

[1] The political science department at Brigham Young University does not offer any graduate degrees.

and quantitative analysis more generally. I have been asked to evaluate the amicus brief submitted on April 23, 2021.

2 The 2020 DAS is an application of statistical inference

In Appendix B of the *Amicus Brief of Data Privacy Experts*, the authors provide an authoritative scholarly reference to define statistical inference:[2]

> "A statistical inference...[is] a statement about statistical populations made from given observations with measured uncertainty. An inference in general is an uncertain conclusion. Two things mark out statistical inferences. First, the information on which they are based is statistical, i.e. consists of observations subject to random fluctuations. Secondly, we explicitly recognize that our conclusion is uncertain, and attempt to measure, as objectively as possible, the uncertainty involved. Fisher uses the expression 'the rigorous measurement of uncertainty.'"

I agree that this is a clear definition of statistical inference. It is also the case that the disclosure avoidance system (DAS) used by the Census Bureau in 2020 is an application of statistical inference according to the definition provided above. I will note two reasons that this is the case from the amicus brief as well as other materials provided by the Census Bureau.

A key to statistical inference is the rigorous quantification of uncertainty associated with estimates derived from statistical methods. This applies to the post-processing algorithm used in the DAS procedure. The amicus brief describes the post-processing procedure as one of simply rounding numbers and making sure negative values that arose from the first step of the DAS procedure are moved up to no longer be negative. However, the process is

[2]D. R. Cox, Some Problems Connected with Statistical Inference, 29 Ann. Math. Statist, 357, 357 (1958)

much more complicated than the amicus brief suggests. The number of ways in which these adjustments could be made across millions of summary statistics (the Census Bureau often refers to these as the differentially private histogram) is enormous, and deciding how exactly to make these adjustment is a computationally large and difficult problem. The May 22, 2020 PowerPoint presentation of Philip Leclerc discusses the uncertainty involved in this process and details efforts that have been made by the Census Bureau to adjust or incorporate new data into the statistical models that determine where such adjustments occur. For example, slides 18 and 19 discuss the use of "statistical tests in post-processing" and the incorporation of statistical estimates from "prior census releases" in the DAS process. These data are then used in a variety of statistical models to determine the optimal adjustments to make to the data after differential privacy has been applied. The PowerPoint further discusses two such common statistical models, ordinary least squares (OLS) and non-negative least squares (NNLS). These procedures are used to minimize error and estimate statistical parameters that are then used to determine where and how to make post-processing adjustments. Of course, if different information were sampled and incorporated into the post-processing algorithm, then different parameters yielding different adjustments would result. Given this, these adjustments and refinements naturally come with uncertainty regarding the particular choice of data to incorporate into the post-processing algorithm and procedure, and this uncertainty translates directly into the parameters that are estimated via these statistical methods.

Another key part of the definition of statistical inference is quantifying measures of uncertainty. For example, the definition above states, "we explicitly recognize that our conclusion is uncertain, and attempt to measure, as objectively as possible, the uncertainty involved." One way in which uncertainty is quantified in statistical inference is through the measurement of confidence intervals. The amicus brief makes several references to the calculation of confidence intervals that could accompany the differentially privatized data (see, for example, pg. 21). In other words, these confidence intervals would be measures of

the uncertainty associated with the noise that has been added to the enumerated values.

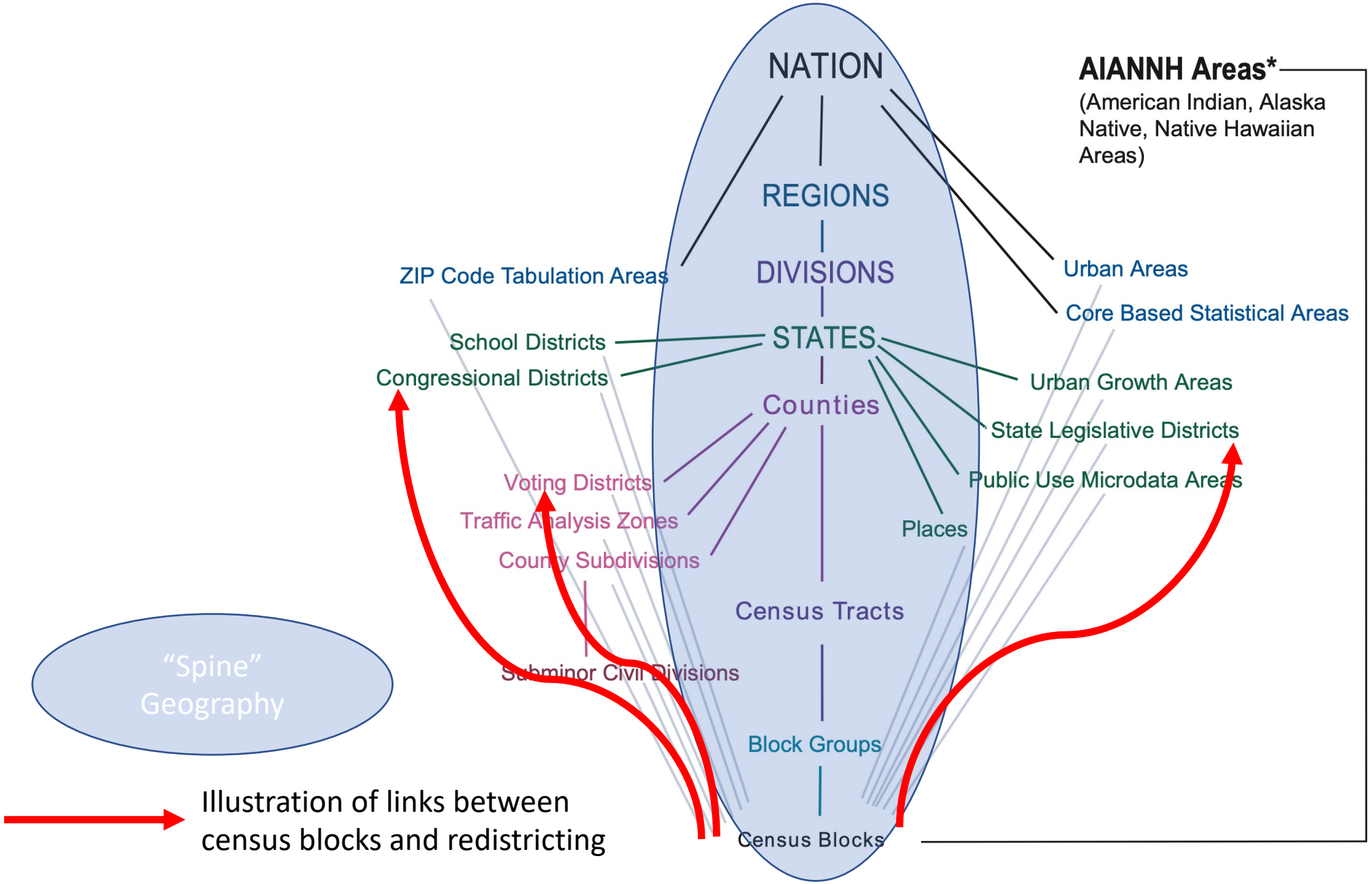
I, Michael Barber, am being compensated for my time in preparing this report at an hourly rate of $400/hour. My compensation is in no way contingent on the conclusions reached as a result of my analysis.

Michael Barber

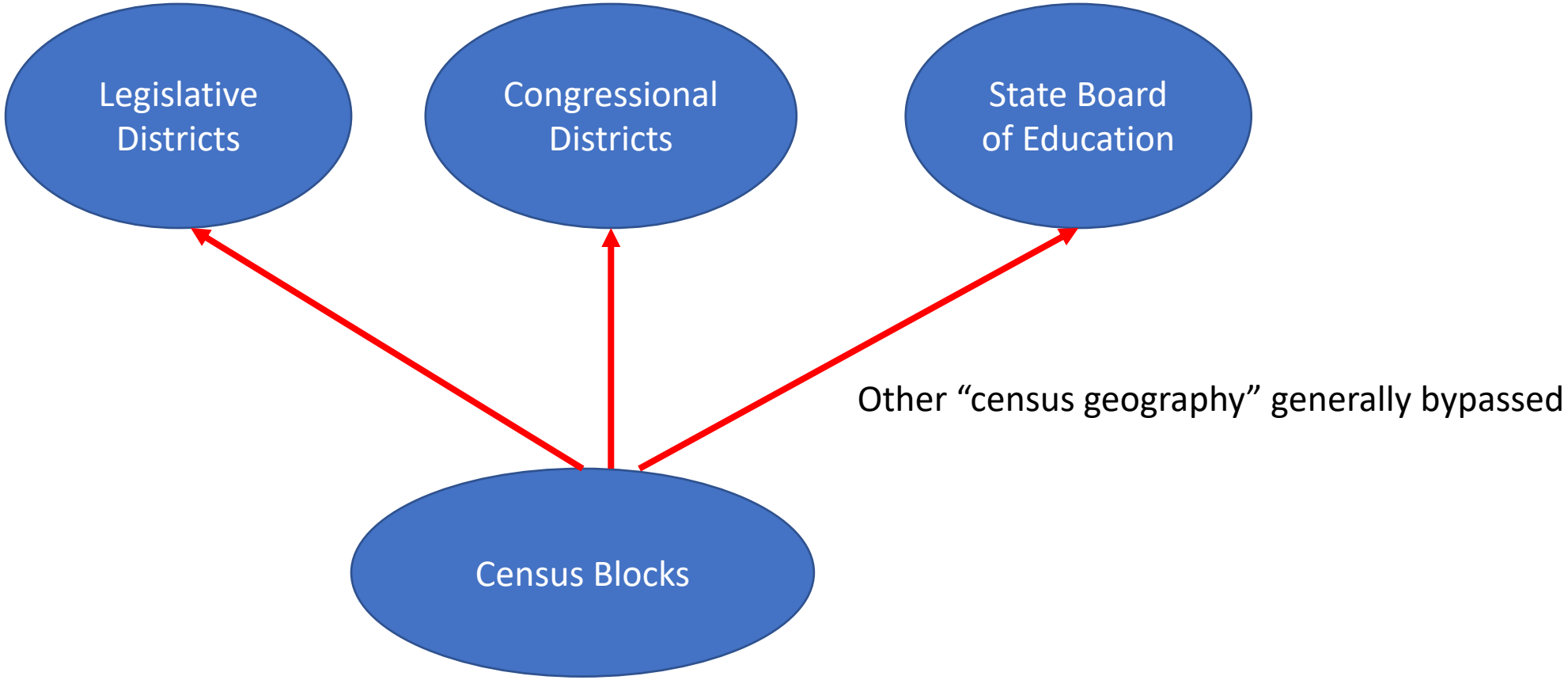
April 26, 2021

DEMONSTRATIVE EXHIBIT 1

https://www2.census.gov/geo/pdfs/reference/geodiagram.pdf (original graphic available)



Building Blocks of Congressional, Legislative and other districts



Building Blocks of Redistricting

- Census Blocks
  - Basic building blocks of geography
    - Water
    - Roads
    - Bridges
    - Mountains
    - City blocks
    - Highways
  - The most basic level of “spine” geography
- Population Data Available When Map Drawing
  - PL94 data
    - Historically – Total population by race, 18+ population by race, Hispanic / Not Hispanic by race for total population, Hispanic / Not Hispanic by race for 18+ population
      - 2020 is adding Group Quarters Population Type, Housing occupancy Status
    - Sometimes computer programs / algorithms are used to break political data down to block level (results in rough estimates)
  - 2010 Data – “average” populated block in Alabama has about 35 people
- Most states, including Alabama draw new Congressional and state legislative districts from the blocks

Federal Requirements for Map Drawers

- Federal Requirements
  - Equal population
    - Article I, Section 2 (Congressional)
    - 14th Amendment (range of 10% is a rebuttable presumption, generally understood at +/- 5% from "ideal," but could also be -3% to +7%)
  - Voting Rights Act
    - Section 2
      - Applies nationwide to every "representational body" election
      - Requires analysis of racially polarized voting (based on precinct election results and racial composition of precincts calculated from underlying census blocks)
      - Requires assessment that "majority of minority voters" reside in an area (based on racial composition generally tabulated at the block, VTD, MCD, Census Place or County level depending on where lines are drawn)
    - Section 5
      - Applies to any jurisdiction "bailed in" under Section 3
      - May apply again this decade if Congress adopts a new "coverage formula"

Racial Block Voting Analysis in Federal Law

- VTD / Precinct
  - VTDs are generally comprised of census blocks established in the calendar year ending in 8
  - Precincts are the lowest level for which “political data” or election outcomes can be accurately measured – election results are reported at precinct level
  - Precincts are the basic unit for racially polarized voting analysis under the Voting Rights Act used in Section 2 and Section 5 compliance and enforcement
  - Racial composition of precincts can be determined by adding up PL94 data from blocks that compose each precinct
  - Precincts are often rebuilt by local officials after each census to smooth out or equalize population / registered voters and align precincts with new Congressional, legislative and other districts
    - Avoids “split precincts” with more than one ballot style
  - Alabama is not required to draw from precincts – and has in the past drawn at block level
  - VTDs are not “spine” geography – and generally shift boundaries for each census
    - Precincts can shift boundaries for any election in between Census

Alabama State Requirements for Map Drawers

- State Requirements for legislative districts
  - Once per decade after the decennial Census (State constitution)
    - State Constitution generally requires use of federal decennial census
    - Required in the first session after the taking of the census
  - Districts must be “contiguous” (for state Senate per state constitution)
- 2011 Guidelines adopted by the legislature for state legislative, congressional and state board of education districts:
  - Federal law compliance (equal population, VRA)
  - 1% population deviation
  - Single member districts
  - Contiguous territory
  - No “subordination” of race-neutral redistricting
  - Draw based on total population
  - Preserving county level political subdivisions
  - Compact districts
  - Avoid contests between incumbents
  - Preserve communities of interest
- 2020 Guidelines have not been adopted

Alabama's "small area" populations & the Bureau's latest assurance of "accuracy"

- Census says that "off spine" geographies "as small as 500 people" will be 99.5% confident that their "largest ethnic group" will be +/- 5% of the actual population in the next round of DP demonstration data
- Professor Andy Beveridge's analysis of "spine" and "off spine" populations less than 500 in Alabama in 2010 Census:
  - 168 of 583 "census places" in Alabama have less than 500 people ("Off spine")
    - About 29% of "census places" in Alabama
  - 225 of the 3432 "block groups" in Alabama have less than 500 people ("spine")
    - Generally not used in map drawing
  - 421 of the 1988 VTDs in Alabama have less than 500 people ("Off spine")
    - More than 20% of VTDs in Alabama will NOT have any assurance from Census as to accuracy of population for largest ethnic group
    - VTDs are set to be re-drawn immediately after release of the census data – so accuracy as to existing VTD population is essentially useless
  - 134,916 of the 135,439 population blocks in Alabama have less than 500 people ("spine")
    - Only 523 census block will have any assurance of accuracy of population as to accuracy of population for largest ethnic group (.38% of populated Census blocks)
  - For each of these "small" areas, Census to date provides no information about the "confidence in variation" that DP injects in the latest demonstration data
  - Source: https://www.socialexplorer.com/blog/post/sixteen-states-sue-to-block-census-bureau-data-privacy-method-11411

Examples of Alabama Counties that Must Contain More than One District

Ideal Populations:

- Congressional = 635,300
- State Senate = 127,060
- State House = 42,353

County	Population	Congressional	Senate	House
Jefferson	658,466	1+	5+	15+
Mobile	412,992		3+	9+
Madison	334,811		2+	7+
Montgomery	229,363		1+	5+
Shelby	195,085		1+	4+
Tuscaloosa	194,656		1+	4+

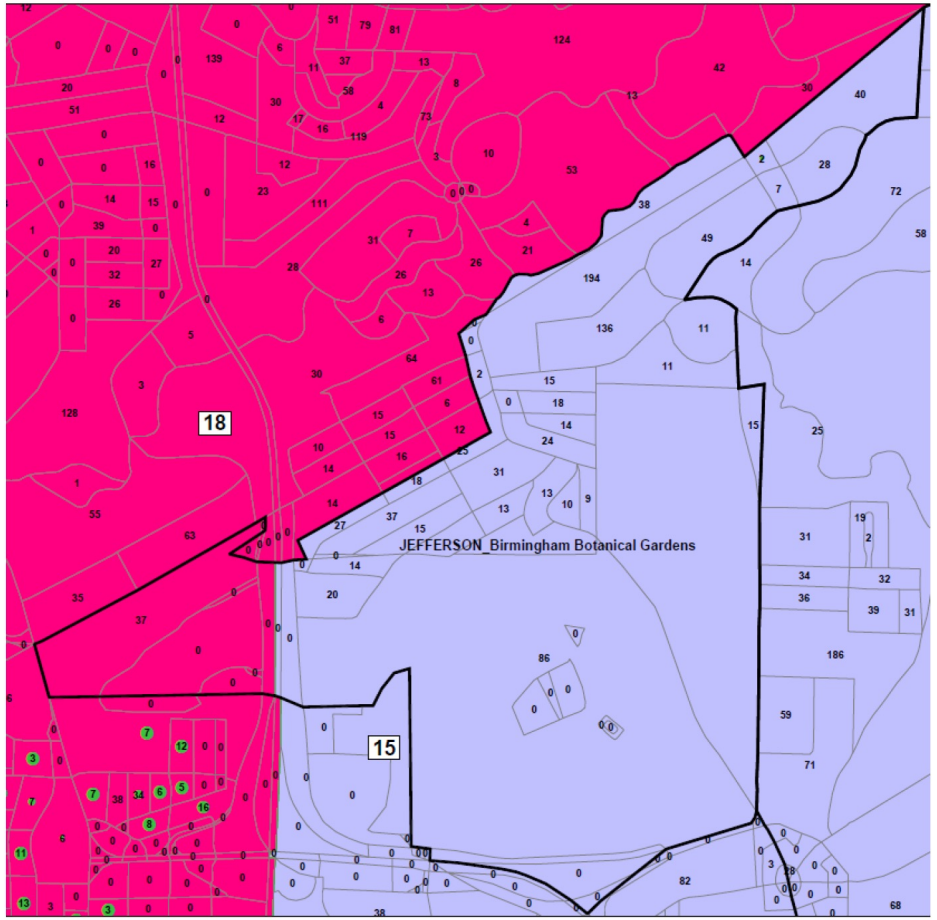
Block Level Data is Critical Component of Judicial Opinions in Alabama

- Alabama Legislative Black Caucus v. Alabama, 989 F.Supp2d 1227 (2013)
  - Precincts split while drawing at the block level (at 1277)
  - Court noted that accurate racial data was available at the block level (at 1319)
  - Court noted that "political" data only available at the precinct level (at 1319)
  - Court reviewed racial characteristics for legislative districts focused on less than 1% BVAP changes in house districts and senate districts (at 1320-1321 at fn 16 and 17)
  - Court highlighted BVAP changes ranging from -.82% to +9.76% (at 1321 at fn 18 and 19)

Block Level Data is Critical Component of Judicial Opinions in Alabama

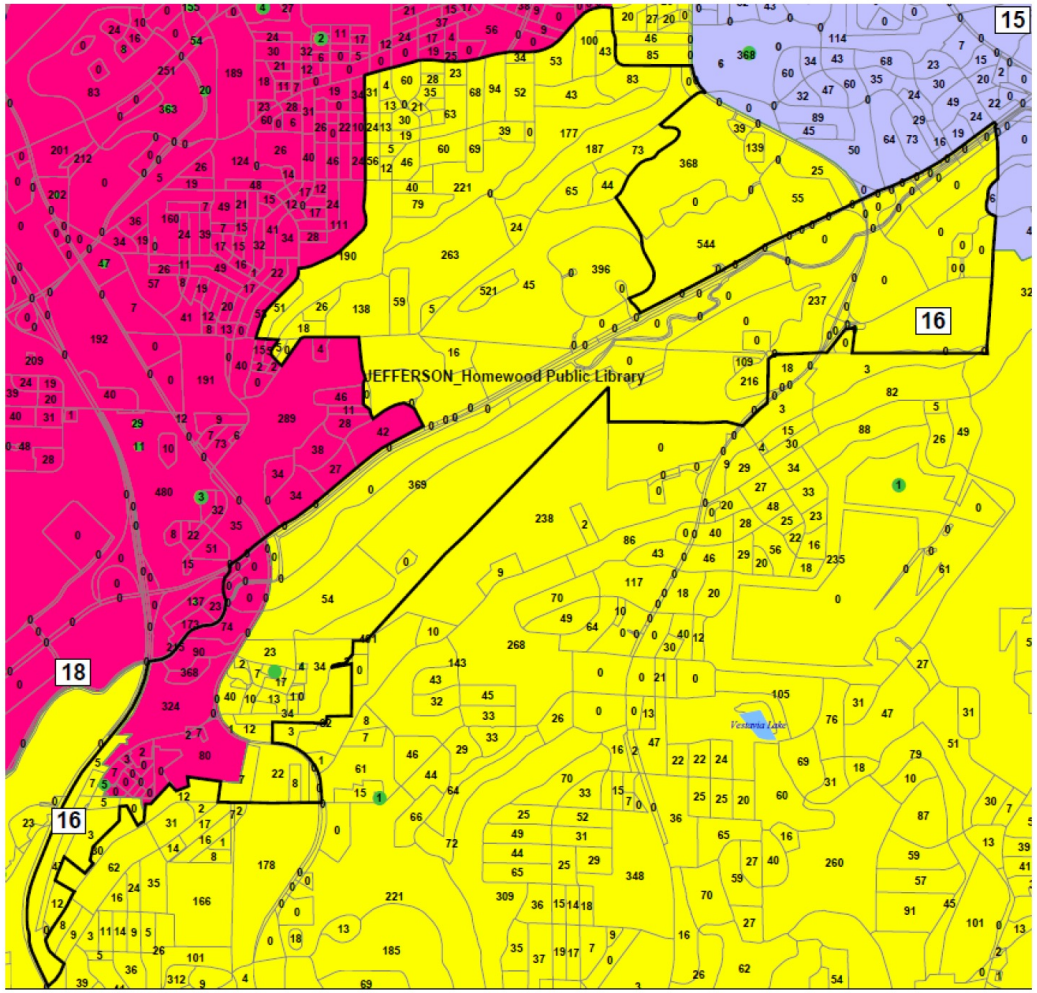
- Alabama Legislative Black Caucus v. Alabama, 231 F.Supp.3d 1026 (2017)
  - Noting availability of political data at precinct level, and race data to block level (at 1038)
  - Discussion throughout of precincts split and examining block level race data of the split precincts
  - Maps with block level analysis of race easily exceeded 200 maps in the 457 pages of majority opinion authored by now Chief Judge Pryor
  - 5 maps with block level analysis by race appeared in the 170 page concurring and dissenting opinion

Sample of Maps included in 2017 Alabama Legislative Black Caucus Decision



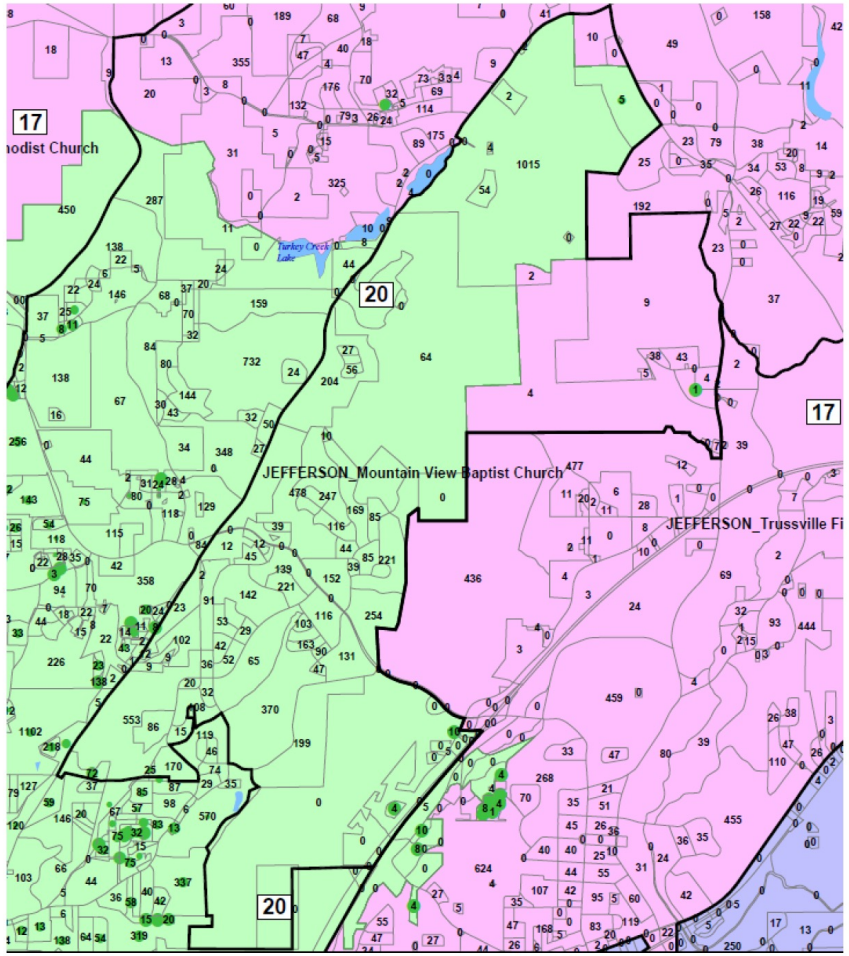
(APSX 317). In Mountain Brook City Hall precinct, the drafters split a precinct of exclusively majority-white blocks between Districts 18 and 15.

Homewood Public Library Precinct in Act 603

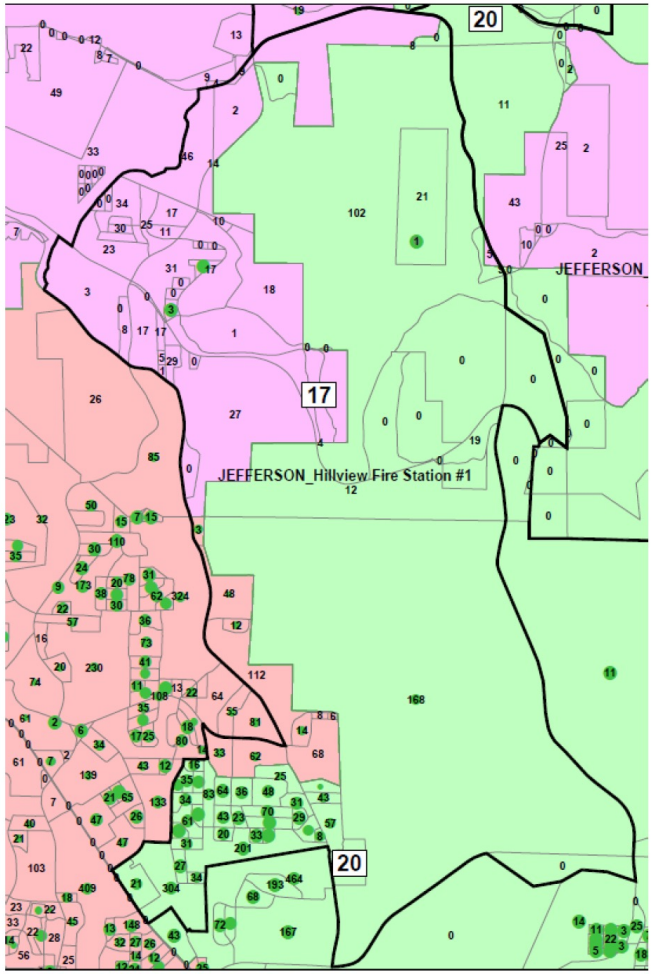


Sample of Maps included in 2017 Alabama Legislative Black Caucus Decision (cont.)

Mountain View Baptist Church Precinct in Act 603

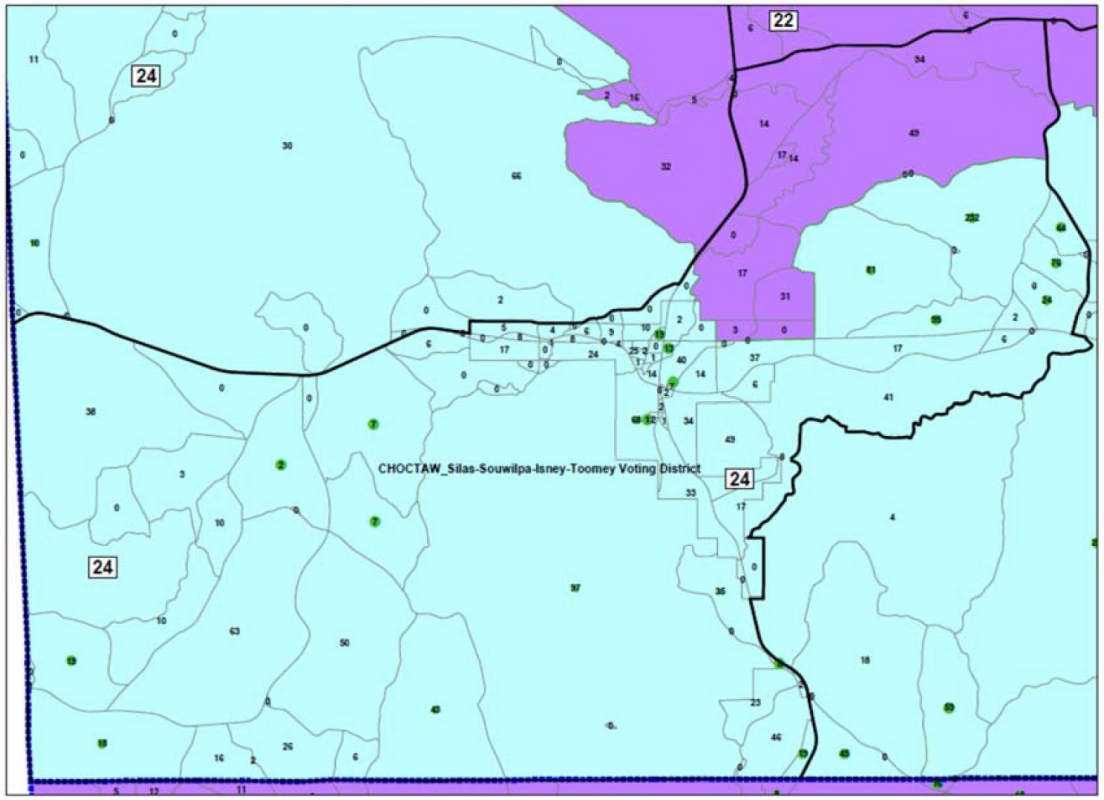


Hillview Fire Station #1 Precinct in Act 603



Sample of Maps included in 2017 Alabama Legislative Black Caucus Dissent

APSX291_Act_603_CHOCTAW_Silas-Souwilpa-Isney-Toomey Voting District



APSX108_JEFFERSON_Gardendale Civic Center