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

STATE OF ALABAMA, *et al.*, Plaintiffs, v. UNITED STATES DEPARTMENT OF COMMERCE, *et al.*, Defendants.	Case No. 3:21-CV-211-RAH-ECM-KCN

DECLARATION OF JOHN M. ABOWD

I, John M. Abowd, make the following Declaration pursuant to 28 U.S.C. § 1746, and declare that under penalty of perjury the following is true and correct to the best of my knowledge:

BACKGROUND

1. I am the Chief Scientist and Associate Director for Research and Methodology at the United States Census Bureau. I have served in this capacity since June 2016. My statements in this declaration are based on my personal knowledge or on information supplied to me in the course of my professional responsibilities.
2. I received my Ph.D. in economics from the University of Chicago with specializations in econometrics and labor economics in 1977 (M.A. 1976). My B.A. in economics is from the University of Notre Dame.
3. I have been a university professor since 1976 when I was appointed assistant professor of economics at Princeton University. I was also assistant and associate professor of econometrics and industrial relations at the University of Chicago Graduate School of Business. In 1987, I was appointed associate professor of industrial and labor relations with indefinite tenure at Cornell University where I am currently the Edmund Ezra Day Professor. I am on unpaid leave from Cornell University to work in my current position at the Census Bureau as part of the Career Senior Executive Service.
4. I am a member and fellow of the American Association for the Advancement of Science, American Statistical Association, Econometric Society, and Society of Labor Economists (president 2014). I am an elected member of the International Statistical Institute. I am also a member of the American Economic Association, International Association for Official Statistics, National Association for Business Economists, American Association for Public Opinion Research, Association for Computing Machinery, and American Association of Wine Economists. I regularly attend and present papers at the meetings of these organizations.

5. I have served on the American Economic Association Committee on Economic Statistics. I have also served on the National Academy of Sciences Committee on National Statistics, the Conference on Research in Income and Wealth Executive Committee, and the Bureau of Labor Statistics Technical Advisory Board for the National Longitudinal Surveys (chair: 1999-2001).
6. I have worked with the Census Bureau since 1998, when the Census Bureau and Cornell University entered into the first of a sequence of Intergovernmental Personnel Act agreements and other contracts. Under those agreements, I served continuously as Distinguished Senior Research Fellow at the Census Bureau until I assumed my current position as Chief Scientist in 2016, under a new Intergovernmental Personnel Act contract. Since March 29, 2020, I have been in the Associate Director position at the Census Bureau as a Career Senior Executive Service employee.
7. From 2011 until I assumed my position as Chief Scientist at the Census Bureau in 2016, I was the lead Principal Investigator of the Cornell University node of the NSF-Census Research Network, one of eight such nodes that worked collaboratively with the Census Bureau and other federal statistical agencies to identify important theoretical and applied research projects of direct programmatic importance to the agencies. The Cornell node produced the fundamental science explaining the distinct roles of statistical policymakers and computer scientists in the design and implementation of differential privacy systems at statistical agencies.
8. I have published more than 100 scholarly books, monographs, and articles in the disciplines of economics, econometrics, statistics, computer science, and information science. I have been the principal investigator or co-principal investigator on 35 sponsored research projects. I was a founding editor of the Journal of Privacy and Confidentiality—an interdisciplinary journal, and I continue to serve as an editor and on the governance board. My full professional resume is attached to this report as Appendix A.

9. I have worked on and managed Census Bureau projects that were precursors to the Census Bureau's current program to implement differential privacy for the 2020 Census of Population and Housing. I was one of three senior researchers who founded the Longitudinal Employer-Household Dynamics (LEHD) program at the Census Bureau, which is generally acknowledged as the Census Bureau's first 21st Century data product: built to the specifications of local labor market specialists without additional survey burden, and published beginning in 2001 using state-of-the-art confidentiality protection via noise infusion. This program produces detailed public-use statistical data on the characteristics of workers and employers in local labor markets using large-scale linked administrative, census, and survey data from many different sources. In 2008, my work with LEHD led to the first production implementation worldwide of differential privacy as part of a product of the LEHD program called OnTheMap. The LEHD program also implemented other prototype systems to protect confidential information, including allowing the public to access synthetic micro-data confirmed via direct analysis of the confidential data on validation servers. A differentially private version of this system is under development at the Census Bureau but not for use with the 2020 Census.

IMPORTANCE OF CONFIDENTIALITY

10. Though participation in the census is mandatory under 13 U.S. Code § 221, in practice, the Census Bureau must rely on the voluntary participation of each household in order to conduct a complete enumeration.
11. One of the most significant barriers to conducting a complete and accurate enumeration are individuals' concerns about the confidentiality of census data. The Census Bureau's pre-2020 Census research showed that 28% of respondents were "extremely concerned" or "very concerned" and a further 25% were "somewhat concerned"

about the confidentiality of their census responses.[1] These concerns are even more pronounced in minority populations and represent a major operational challenge to enumerating traditionally hard-to-count populations.[2]

12. To secure voluntary participation, Congress first established confidentiality protections for individual census responses in the Census Act of 1879. These confidentiality protections were later expanded and codified in 13 U.S. Code §§ 8(b) & 9, which prohibits the Census Bureau from releasing "any publication whereby the data furnished by any particular establishment or individual under this title can be identified[,]" and allows the Secretary to provide aggregate statistics so long as those data "do not disclose the information reported by, or on behalf of, any particular respondent[.]" Title III of the Foundations for Evidence Based Policymaking Act of 2018 also requires statistical agencies to "protect the trust of information providers by ensuring the confidentiality and exclusive statistical use of their responses."[3]
13. The broader scientific community generally concurs about the importance of rigorous protection of confidentiality by statistical agencies. For example, the National Academy of Sciences' definitive guidebook for federal statistical agencies states "Because virtually every person, household, business, state or local government, and organization is the subject of some federal statistics, public trust is essential for the continued effectiveness of federal statistical agencies. Individuals and entities providing data di-

[1] U.S. Census Bureau (2019) "2020 Census Barriers, Attitudes, and Motivators Study Survey Report" https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-cbams-study-survey.pdf, p.38-39.

[2] Ibid, p.39-42.

[3] Title III of the Foundations for Evidence Based Policymaking Act of 2018, § 3563.

rectly or indirectly to federal statistical agencies must trust that the agencies will appropriately handle and protect their information."[4] The report also notes that respondents expect statistical agencies not to "release or publish their information in identifiable form."[5] The National Academies also broadly exhort statistical agencies to "continually seek to improve and innovate their processes, methods, and statistical products to better measure an ever-changing world."[6]

14. The Census Bureau enjoys higher self-response rates than private survey companies in large part because the public generally trusts the Census Bureau to keep its data safe. The Census Bureau makes extensive outreach efforts to assure respondents and other data providers about the Bureau's commitment to protection of confidential data. The criminal fines and imprisonment penalties that Census Bureau employees would face by unlawfully disclosing respondent information are frequently cited by the Census Bureau in these outreach efforts.[7]

15. This trust in the Census Bureau is particularly important for the decennial census, given the "civic ceremony" aspect of the census, akin to the civic ceremony aspect of elections and voting. The decennial census is an exercise where the nation comes together every ten years, under a strict promise of confidentiality, to provide information to help govern our nation. Were the Census Bureau to expose confidential information, there is no doubt that self-response rates would drop, increasing survey

---

[4] National Academies of Sciences, Engineering, and Medicine 2021. Principles and Practices for a Federal Statistical Agency: Seventh Edition. Washington, DC: The National Academies Press. https://doi.org/10.17226/25885, p. 37-38.

[5] Ibid., p.38.

[6] Ibid., p.4.

[7] https://www.census.gov/content/dam/Census/library/fact-sheets/2019/comm/2020-confidentiality-factsheet.pdf.

cost across programs by increasing in-person follow up, and decreasing the quality of the census overall.

PRIVACY PROTECTION AT THE CENSUS BUREAU

16. Protecting privacy is at the core of the Census Bureau's mission. Our privacy promise to respondents is key to promoting response to our censuses and surveys. The Census Bureau—at the crux of its dual mandate to publish only statistical summaries and to protect the confidentiality of respondent data—is balancing the preferences of data users and data providers. An optimal choice must account for the preferences of data users and protect the data the American people entrust the Census Bureau with keeping safe. [8]
17. Data collected from the decennial census support a wide array of critical government and societal functions at the federal, state, tribal, and local levels. In addition to apportioning seats in the U.S. House of Representatives and supporting the redistricting of those seats, census data also support the allocation of over $675 billion in federal

[8] "Official Statistics at the Crossroads: Data Quality and Access in an Era of Heightened Privacy Risk," *The Survey Statistician*, 2021, Vol. 83, 23-26 (available at Survey_Statistician_2021_January_N83_03.pdf (isi-iass.org)). The paper is based on talks that I gave in 2019 to the Committee on National Statistics and the Joint Statistical Meetings. It summarizes the research in Abowd, J.M. and I. Schmutte "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices," *American Economic Review,* Vol. 109, No. 1 (January 2019):171-202, DOI:10.1257/aer.20170627.

funding each year based on population counts, geography, and demographic characteristics.[9] Census data also support important public and private sector decision-making at the federal, state, tribal, and local levels, and serve as benchmark statistics for other important surveys and data collections throughout the decade.[10]

18. The Census Bureau publishes an enormous number of statistics calculated from its collected data. Following the 2010 Census, for example, the Census Bureau published over 150 billion independent statistics about the characteristics of the 308,745,538 persons in the resident population that were enumerated in the census. To serve their intended governmental and societal uses, the majority of these statistics needed to be published at very fine levels of detail and with geographic precision often down to the individual census tract or block.

19. While it would be quite difficult from any single one of those published statistics to ascertain the identity of any individual census respondent or the contents of that respondent's census response, the volume and detail of information published by the Census Bureau, taken together, pose a serious challenge for protecting the privacy and confidentiality of census data. Combining information from multiple published statistics or tables can make it easy to pick out those individuals in a particular geographic area whose characteristics differ from those of the rest of their neighbors. These individuals, who have unique combinations of the demographic characteristics

[9] Hotchkiss, M., & Phelan, J. (2017). Uses of Census Bureau data in federal funds distribution: A new design for the 21st century. United States Census Bureau. https://www2.census.gov/programs-surveys/decennial/2020/program-management/working-papers/Uses-of-Census-Bureau-Data-in-Federal-Funds-Distribution.pdf.

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

reported in statistical summaries, are known as "population uniques" and their records have traditionally been the target of the mechanisms that the Census Bureau uses to protect confidentiality in its data publications.

20. Traditional statistical disclosure limitation methods,[11] like those used in 2010 census, cannot defend against modern challenges posed by enormous cloud computing capacity and sophisticated software libraries. That does not mean traditional statistical disclosure limitation methods usually fail—they usually do not fail. But as computer scientists bring their expertise from the field of cryptography to the field of safe data publication, they have exposed significant vulnerabilities in traditional privacy methods. The Census Bureau's own internal analysis, for example, confirmed that a modern database reconstruction-abetted re-identification attack can reliably match a large number of 2010 census responses to the names of those respondents—a vulnerability that exposed information of *at least* 52 million Americans and potentially up to 179 million Americans.[12] To defend against this known vulnerability, the Census Bureau explored different confidentiality methods that explicitly defend against database reconstruction attacks and concluded that the best tool to protect against this modern attack while also preserving the accuracy and usability of data products comes from the body of scientific work called "differential privacy."

THE HISTORY OF INNOVATION IN THE DECENNIAL CENSUS

21. The decennial census, known officially as the *Decennial Census of Population and Housing,* is the flagship statistical product of the U.S. Census Bureau. Though the Census

[11] The technical field that addresses confidentiality is known as "statistical disclosure limitation." At the Census Bureau, it is known as "disclosure avoidance." It is also called "statistical disclosure control" by some statisticians and "privacy-preserving data analysis" by some computer scientists.

[12] See Appendix B for a summary of the Census Bureau's simulated reconstruction and re-identification attacks.

Bureau conducts hundreds of surveys every year, the once-every-decade enumeration of the population of the United States, mandated by Article I, Section 2 of the U.S. Constitution, is the single largest and most complex data collection regularly conducted by the United States government. Since the very first U.S. census in 1790, the collection, processing, and dissemination of census data have posed unique challenges and have required the Census Bureau to improve its operations every decade.

22. The challenges faced by the Census Bureau have led to remarkable innovations. Herman Hollerith's electric tabulation machine, developed for the 1890 Census, revolutionized the field of data processing and led Hollerith to form the company that eventually became IBM.[13] To conduct the 1950 Census, the Census Bureau commissioned the development of the first successful civilian digital computer, UNIVAC I.[14] With each passing decade, the Census Bureau develops, tests, and deploys innovations to its statistical methods, field data collection methods, and data processing operations.

23. That spirit of innovation includes the Census Bureau's more recent implementation of cutting-edge privacy protections. Prior to the 1990 Census, the primary mechanism that the Census Bureau employed to protect the confidentiality of individual census responses was to withhold publication of (or "suppress") any table that did not meet certain household, population, or demographic characteristic thresholds. The 1970 Census, for example, suppressed tables reflecting fewer than five households, and would only publish tables of demographic characteristics cross-tabulated by race if

[13] https://www.census.gov/history/www/census_then_now/notable_alumni/herman_hollerith.html.

[14] https://www.census.gov/history/www/innovations/technology/univac_i.html.

there were at least five individuals in each reported race category.[15] These suppression routines helped to protect privacy by reducing the detail of data published about individuals who were relatively unique within their communities. By the 1990 Census, however, the Census Bureau transitioned away from suppression methodologies for two reasons: first, data users were dissatisfied with missing details caused by suppression and second, the Bureau realized that the suppression routines it had been using were insufficient to fully protect against re-identification.[16]

24. For the 1990 Census, the Bureau began using a technique known as noise infusion to safeguard respondent confidentiality. Noise infusion helps to protect the confidentiality of published data by introducing controlled amounts of error or "noise" into the data. The goal of noise infusion is to preserve the overall statistical validity of the resulting data while introducing enough uncertainty that attackers would not have any reasonable degree of certainty that they had isolated data for any particular respondent. The noise infusion used in 1990 was a very simple form of data swapping between paired households in a geographic area with similar attributes, and for small block groups the Census Bureau replaced the collected characteristics of households with imputed characteristics.[17]

---

[15] Zeisset, P. (1978), "Suppression vs. Random Rounding: Disclosure Avoidance Alternatives for the 1980 Census," https://www.census.gov/content/dam/Census/library/working-papers/1978/adrm/Suppression%20vs.%20Random%20Rounding%20Disclosure-Avoidance%20Alternatives%20for%20the%201980%20Census.pdf.

[16] McKenna, L. (2018), "Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing," https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf, p.6.

[17] Ibid., p. 6-7. An "imputed characteristic" is the prediction of a statistical model used in place of a missing characteristic, when used in standard editing procedures, or in place of a collected characteristic, when used for confidentiality protection.

25. For the 2000 and 2010 censuses, the Census Bureau began to infuse noise using a more advanced "data swapping" method. The Census Bureau first identified households most vulnerable to re-identification—especially households on smaller-population blocks whose residents had differing demographic characteristics from the remainder of their block. While every non-imputed[18] household record in the Census Edited File (CEF) had a chance of being selected for data swapping, records for more vulnerable households (typically those on low-population blocks) were selected with greater probability. Then, the records for all members of those selected households were exchanged with the records of households in nearby geographic areas that matched on key characteristics. For the 2000 and 2010 censuses, those key matching characteristics were (1) the whole number of persons in the household, and (2) the whole number of persons aged 18 or older in the household. These swapping criteria resulted in the total population and total voting age population for each block being held "invariant"—that is, while noise was added to all remaining characteristics, no noise was added to the block-level total population or block-level voting age population counts.[19] *The selection and application of these particular invariants is not an innate feature of data swapping; invariants are implementation parameters that can be applied to (or removed from) any counted characteristic under any noise infusion methodology.*

[18] When a respondent household provides only a count of the number of persons living at that address or when the housing unit population count is itself imputed, the Census Bureau imputes all characteristics: sex, age, race, ethnicity, and relationship to others in the household. Such persons are called "whole-person census imputations" in technical documentation. When a household consists entirely of whole-person census imputation records, it is called an "imputed" household. A "non-imputed" household contains at least one person whose characteristics were collected on the census form for the household.

[19] Ibid. p. 8-10.

THE PRIVACY PROTECTIONS USED FOR THE 2010 CENSUS ARE NO LONGER SUFFICIENT

26. While the Census Bureau's confidentiality methodologies for the 2000 and 2010 censuses were considered sufficient at the time, advances in technology in the years since have reduced the confidentiality protection provided by data swapping.

27. Disclosure avoidance has been a recognized branch of statistics since the 1970s, but it has only been since the late 1990s that it has evolved into a distinct scientific field of study in both statistics and computer science. Prof. Latanya Sweeney's 1997 revelation that she had re-identified then Massachusetts Governor William Weld's medical records in a purportedly "deidentified" public database[20] prompted the Census Bureau and many other statistical agencies to re-examine the efficacy of their disclosure avoidance techniques.

28. *Re-identification attacks.* Prior to 2016, disclosure risk assessments usually focused on assessing the vulnerability of microdata releases (data products that contain individual records for all or some of the data subjects deidentified by removing names and addresses), rather than the rules used for aggregated data releases (data compiled and aggregated into tables). Simulated "re-identification attacks" analyze the risk that an external attacker could use individuals' characteristics that are included on a published microdata file (e.g., location, age, and sex) and link those records to a third-party data source (e.g., commercial data or voter registration lists) that contains those characteristics along with the individuals' names and addresses. The resulting rates of "putative" (suspected) and confirmed linkages show the overall degree of vulnerability of the data. If those linkage rates are deemed too large, then additional disclosure avoidance is necessary to mitigate the disclosure risk.

[20] Sweeney, L. (2002). "k-anonymity: a model for protecting privacy." International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5); 557-570, also recounted in Ohm, P. (2009) "Broken promises of privacy: Responding to the surprising failure of anonymization." UCLA l. Rev. 57: 1701.

29. The general problem with relying exclusively on re-identification studies to assess disclosure risk is that they can only provide a "best-case" approximation of the underlying disclosure risk of the data. If a real attacker has access to more sophisticated tools (e.g., optimization algorithms or computing power) or to higher quality external data (e.g., with better age and address information) than the tools or data used in the simulated attack, then the real disclosure risk will be substantially higher than what is estimated via the study. This limitation is particularly vexing for statistical agencies that must rely on a "release and forget" approach to data publication, where disclosure avoidance safeguards must be selected without foreknowledge of the better tools and external data that attackers may have at their disposal after the data are published.
30. Re-identification studies also underestimate the risk from releasing aggregated data. The Census Bureau has long relied on re-identification studies to assess the disclosure risk of its microdata releases, but the majority of Census Bureau data products are aggregated data releases. Over the past decade, aggregated data releases have become increasingly vulnerable to sophisticated "reconstruction attacks" that have emerged as computing power has improved and gotten substantially cheaper.
31. *Reconstruction attacks.* The theory behind a "reconstruction attack" is that the release of *any* statistic calculated from a confidential data source will reveal a potentially trivial, but non-zero, amount of confidential information.[21] As a consequence, if an attacker has access to enough aggregated data with sufficient detail and precision, then the attacker may be able to leverage information from each statistic in the aggregated data to reconstruct the individual-level records that were used to generate the published tables. This process is known as a "reconstruction attack," and it adds a new

[21] Dinur, I. and Nissim, K. (2003) "Revealing Information while Preserving Privacy" PODS, June 9-12, San Diego, CA. https://doi.org/10.1145/773153.773173.

degree of disclosure vulnerability against which statistical agencies must defend. While the statistical and computer science communities have been aware of this vulnerability since 2003, only over the last few years have computing power and the sophisticated numerical optimization software necessary to perform these types of reconstructions advanced enough to permit reconstruction attacks at any significant scale.

32. The risk of reconstruction and re-identification attacks is real and substantiated. The Census Bureau has been approached by Prof. Sweeney and others who claim that they have identified specific vulnerabilities in our standard disclosure avoidance methodologies.[22] The vulnerabilities in the disclosure avoidance protections for the Census Bureau's Survey of Income and Program Participation (SIPP) identified by Prof. Sweeney led the Census Bureau to immediately implement permanent changes to the disclosure avoidance rules used for SIPP data, including increased noise infusion and delayed reporting of survey participants' major life events.[23]

33. Statistical releases do not all need to be of the same type, or contain the same data fields, to enable re-identification by reconstruction. For example, a 2015 interagency report published by the National Institute of Standards and Technology (NIST) written by my colleague Simson Garfinkel provided examples of using disparate data sets to reconstruct hidden underlying data.[24] Some of these examples are quoted here:

---

[22] McKenna, L. (2019b). "U.S. Census Bureau Reidentification Studies," available at https://www.census.gov/library/working-papers/2019/adrm/2019-04-ReidentificationStudies.html.

[23] McKenna, L. (2019b). p. 2-3.

[24] Garfinkel, S. (2015) "De-Identification of Personal Information," National Institute of Standards and Technology, available at http://dx.doi.org/10.6028/NIST.IR.8053 at 26-27.

34. "*The Netflix Prize:* Narayanan and Shmatikov showed in 2008 that in many cases the set of movies that a person had watched could be used as an identifier.[25] Netflix had released a dataset of movies that some of its customers had watched and ranked as part of its "Netflix Prize" competition. Although there was [sic] no direct identifiers in the dataset, the researchers showed that a set of movies watched (especially less popular films, such as cult classics and foreign films) could frequently be used to match a user profile from the Netflix dataset to a single user profile in the Internet Movie Data Base (IMDB), which had not been de-identified and included user names, many of which were real names. The threat scenario is that by rating a few movies on IMDB, a person might inadvertently reveal *all* of the movies that they had watched, since the person's IMDB profile could be linked with the Netflix Prize data."[26] (emphasis in original)

35. "*Credit Card Transactions:* Working with a collection of de-identified credit card transactions from a sample of 1.1 million people from an unnamed country, Montjoye *et al.* showed that four distinct points in space and time were sufficient to specify uniquely 90% of the individuals in their sample.[27] Lowering the geographical resolution and binning transaction values (*e.g.,* reporting a purchase of $14.86 as between $10.00 and $19.99) increased the number of points required."[28]

36. "*Mobility Traces:* Montjoye *et al.* showed that people and vehicles could be identified by their "mobility traces" (a record of locations and times that the person or vehicle

---

[25] Narayanan, A. and Shmatikov V. "Robust De-anonymization of Large Sparse Datasets," *IEEE Symposium on Security and Privacy* (2008): 111-125.

[26] Garfinkel, S. (2015), p. 26-27.

[27] Montjoye, Y-A. et al. "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, 30 (January 2015) Vol 347, Issue 6221.

[28] Garfinkel, S. (2015), p. 27.

visited). In their study, trace data from a sample of 1.5 million individuals was processed, with time values being generalized to the hour and spatial data generalized to the resolution provided by a cell phone system (typically 10-20 city blocks).[29] The researchers found that four randomly chosen observations of an individual putting them at a specific place and time was sufficient to uniquely identify 95% of the data subjects.[30] Space/time points for individuals can be collected from a variety of sources, including purchases with a credit card, a photograph, or Internet usage. A similar study performed by Ma *et al.* found that 30%-50% of individuals could be identified with 10 pieces of side information.[31] The threat scenario is that a person who revealed five place/time pairs (perhaps by sending email from work and home at four times over the course of a month) would make it possible for an attacker to identify his or her entire mobility trace in a publicly released dataset. As above, the attacker would need to know that the target was in the data."[32]

37. The same general principles apply to census data. The difference between census data and the examples above is that census data can be combined in vastly more ways with other information because all the tables published from census data share basic standardized identifiers including location, age, sex, race, ethnicity, and marital status. Even if each of these identifiers is not included in every table, their use and combinations across many different tables creates the disclosure risk. The Census Bureau understood this emerging risk even before the 2010 Census. As field collection for the

[29] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3(1).

[30] Ibid., p. 1-5.

[31] C. Y. T. Ma, D. K. Y. Yau, N. K. Yip and N. S. V. Rao (2013) "Privacy Vulnerability of Published Anonymous Mobility Traces," in IEEE/ACM Transactions on Networking, vol. 21, no. 3, pp. 720-733, June 2013, doi: 10.1109/TNET.2012.2208983.

[32] Garfinkel, S. (2015), p. 27-28.

2010 Census was first beginning, the Census Bureau had already flagged the heightened disclosure risk of releasing detailed block level population data, even with the 2010 Census swapping mechanism in place.[33] After tracking this growing risk of reconstruction and re-identification attacks for several years, the Census Bureau decided in 2015 to establish a new team to comprehensively evaluate the Census Bureau's disclosure avoidance methods to determine if they were sufficient to protect against these disclosure risks.[34]

2010 CENSUS SIMULATED RECONSTRUCTION-ABETTED RE-IDENTIFICATION ATTACK

38. The results from the Census Bureau's 2016-2019 research program on simulated reconstruction-abetted re-identification attack were conclusive, indisputable, and alarming. Appendix B, attached to this declaration, provides an overview of that simulation and the results. The bottom line is that our simulated attack showed that a conservative attack scenario using just 6 billion of the over 150 billion statistics released in 2010 would allow an attacker to accurately re-identify *at least* 52 million 2010 Census respondents (17% of the population) and the attacker would have a high degree of confidence in their results with minimal additional verification or field work. In a more pessimistic scenario, an attacker with access to higher quality commercial name and address data than those used in our simulated attack could accurately re-identify around 179 million Americans or around 58% of the population.

---

[33] During a January 2010 meeting of the Census Bureau's Data Stewardship Executive Policy (DSEP) Committee, the chair of the Disclosure Review Board voiced her concerns about the 2010 Census swapping mechanism's ability to adequately protect future censuses, noting specifically the challenge posed by "continuing to release data at the block level, as block populations continue to decrease (e.g., 40% of blocks in North Dakota have only 1 household in them)" Based on this warning, DSEP decided that "the problem of block population size and disclosure avoidance is real, and that it deserves attention in the context of 2020 planning." DSEP Meeting Record, January 14, 2010. See Appendix C.

[34] DSEP Meeting Record, February 5, 2015. See Appendix D.

39. Emerging attack scenarios and our own internal simulated attacks show that were the Census Bureau to use the disclosure avoidance mechanism implemented for the 2010 Census again for the 2020 Census, the results would be vulnerable to reconstruction and re-identification attacks because of the parameters of the swapping mechanism's 2010 implementation: an overall insufficient level of noise, the invariants preserved without noise, and the geographic and demographic detail of the published summary data. The Census Bureau can no longer rely on the swapping implementation used in 2010 if it is to meet its obligations to protect respondent confidentiality under 13 U.S. Code §§ 8(b) & 9. Protecting against new technology-enabled re-identification attacks, while maintaining the high quality of the decennial census data products, requires the implementation of a disclosure avoidance mechanism that is better able to protect against these new, sophisticated vectors of attack.

DISCLOSURE AVOIDANCE OPTIONS CONSIDERED FOR THE 2020 CENSUS

40. Faced with this compelling mathematical and empirical evidence of the inherent vulnerability of the 2010 Census swapping mechanism to protect against reconstruction-abetted re-identification attacks, the Census Bureau began exploring the available data protection strategies that it could employ for the 2020 Census. The three methods the Census considered were *Enhanced Data Swapping*, *Suppression*, and *Differential Privacy*.

41. The Census Bureau decided that differential privacy was the best tool after analyzing the various options through the lens of economics. Efficiently protecting privacy can be viewed as an economic problem because it involves the allocation of a scarce resource—confidential information—between two competing uses: public data products and privacy protection. If we produce more accuracy, we will have less privacy, and vice versa. And just like in the classic economic example of the trade-off between

producing guns and butter, the tradeoff between privacy and accuracy can be analyzed with a production possibility curve. Our empirical analysis showed that differential privacy offered the most efficient trade-off between privacy and accuracy—our calculations showed that the efficiency of differential privacy dominated traditional methods.[35] In other words, regardless of the level of desired confidentiality, differential privacy will always produce more accurate data than the alternative traditional methods considered by the Census Bureau.

42. *Enhanced Data Swapping.* Enhancing the data swapping mechanism used for the 2010 Census in a manner sufficient to protect against emerging threats like reconstruction attacks would have a significant, detrimental impact on data quality. With an estimated 57% of the population[36] known to be unique at the block level, a swapping mechanism that targets vulnerable households for swapping would require significantly higher rates of swapping than were used in 2010 to protect against a reconstruction attack. Implementing swapping in 2020 would also require abandoning the total population and voting-age population invariants that were used in 2010. There are two technical reasons for this. First, at swap rates sufficient to counter the reconstruction of microdata accurate enough to enable large-scale reidentification, it is impossible to find enough paired households with the same number of persons and adults without searching well outside the neighborhood of the original household. Finding swap pairs was a challenge for some states even at the 2010 swap rate. Second,

[35] See Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. American Economic Review, 109(1), 171-202.

[36] Fifty-seven percent of the 308,745,538 person records in the confidential 2010 Census Edited File, the definitive source for all 2010 Census tabulations, were unique on their block location, sex, age (in years), race (any combination of the 6 OMB-approved race categories, 63 possibilities in all) and Hispanic/Latino ethnicity. This previously confidential statistic was approved for publication with DRB clearance number CBDRB-FY21-DSEP-003.

holding the total and adult populations invariant gives the attacker a huge reconstruction advantage—exact record counts in each block for persons and adults. This advantage vastly improves the accuracy of the reconstructed data. Even a small amount of uncertainty about the block location of an individual greatly expands the variability in the reconstructed microdata effectively reducing the chances of a correct linkage in a re-identification attack. If a block is known to contain exactly seven persons in the confidential data, then every feasible reconstructed version of those data will have exactly seven records in that block, meaning that the block identifier will be correct on every record of every feasible reconstructed database. But if the block population is reported with some random fluctuation around seven, then only by chance will the block identifier be correct in the reconstructed data. Compound this effect over 8,000,000 blocks and the number of feasible reconstructions explodes exponentially. This is what provides the protection against re-identification from the reconstructed data.[37] Internal experiments also confirmed that increasing the swap rate from the level used in 2010 and removing the invariants on block-level population counts (to permit the increased level of swapping and protect against reconstruction attacks) would render the resulting data unusable for most data users.

43. *Suppression.* While the Census Bureau could use suppression to protect from a reconstruction attack, the resulting data would be only available at a very high level of generality. Today's data users, including redistricters, rely on detailed block and tract-level data, which would not be available for many areas if the Census were to return to suppression to protect against modern attacks.

44. *Differential Privacy.* Differential privacy, first developed in 2006, is a framework for quantifying the precise disclosure risk associated with each incremental release from

[37] Garfinkel, S., Abowd, J. M., & Martindale, C. (2018). Understanding Database Reconstruction Attacks on Public Data: These attacks on statistical databases are no longer a theoretical danger. Queue, 16(5), 28-53.

a confidential data source.[38] In turn, this allows an agency like the Census Bureau to quantify the precise amount of statistical noise required to protect privacy. This precision allows the Census to calibrate and allocate precise amounts of statistical noise in a way that protects privacy while maintaining the overall statistical validity of the data.

45. The Census Bureau first began using differential privacy to protect its statistical data products in 2008, with the launch of its OnTheMap tool for employee commuting statistics and its heavily used extension OnTheMap for Emergency Management. In the years since, the Census Bureau has also successfully used differential privacy in a number of other innovative statistical products, such as the Post-Secondary Employment Outcomes and Veteran Employment Outcomes products. Differential privacy is also being used by many of the major technology firms, including Apple[39], Google,[40] Microsoft,[41] and Uber.[42] Other statistical agencies, such as the Statistics of Income Di-

[38] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer, Berlin, Heidelberg.

[39]Differential Privacy Team. (2017). "Learning with Privacy at Scale." Apple Machine Learning Journal, 1(8).

[40]Erlingsson, U., V. Pihur, and A. Korolova. (2014). "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response." Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14, 1054–1067.

[41] Ding, B., J. Kulkarni, and S. Yekhanin. (2017). "Collecting Telemetry Data Privately." Advances in Neural Information Processing Systems 30.

[42] Near, J. (2018) "Differential Privacy at Scale: Uber and Berkeley Collaboration," Enigma 2018 (January) USENIX Assoc. https://www.usenix.org/node/208168.

vision of the Internal Revenue Service, have also begun implementing differential privacy.[43] Internationally, the Australian Bureau of Statistics,[44] the Office of National Statistics in the United Kingdom,[45] and Statistics Canada[46] explicitly recognize the threat from combining multiple statistical tabulations to re-identify respondent information and recommend output noise infusion systems, including differential privacy.

46. Faced with the alarming results of the simulated reconstruction attack, which indicated that the established swapping mechanism resulted in far less disclosure protection than it was intended to provide, and considering the available alternatives, the Census Bureau's Data Stewardship Executive Policy Committee (DSEP)[47] determined

---

[43] Bowen, C. et al. (2020) "A Synthetic Supplemental Public-Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications," (July) Tax Policy Center, The Brookings and Urban Institutes. https://www.urban.org/sites/default/files/publication/102547/a-synthetic-supplemental-public-use-file-of-low-income-information-return-data_2.pdf.

[44] Australian Bureau of Statistics, (2019) "Protecting the Confidentiality of Providers," January 2019, *1504.0 - Methodological News*, https://www.abs.gov.au/ausstats/abs@.nsf/Previousproducts/1504.0Main%20Features9999Jan%202019?opendocument&tabname=Summary&prodno=1504.0&issue=Jan%202019&num=&view=, accessed on March 31, 2021.

[45] United Kingdom Office for National Statistics, (2021) "Policy on Protecting Confidentiality in Tables of Birth and Death Statistics," https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyonprotectingconfidentialityintablesofbirthanddeathstatistics#annex-a-understanding-the-legal-and-policy-framework, accessed on March 31, 2021.

[46] Statistics Canada, (2021) "A Brief Survey of Privacy Preserving Technologies," March 2021, *Data Science Network for the Federal Public Service*, https://www.statcan.gc.ca/eng/data-science/network/privacy-preserving, accessed on March 31, 2021.

[47] The Data Stewardship Executive Policy Committee (DSEP) is a committee chaired by the Deputy Director/Chief Operating Officer and composed of career senior executives with expertise in confidentiality practice, the uses of Census Bureau data, and policy.

that the Census Bureau should proceed with the deployment and testing of differential privacy for use in the 2020 Census given its obligations to produce high quality statistics from the decennial census while also protecting the confidentiality of respondents' census records under 13 U.S. Code §§ 8(b) & 9.[48]

47. The best disclosure avoidance option that offers a solution capable of addressing the new risks of reconstruction-abetted re-identification attacks, while preserving the fitness-for-use of the resulting data for the important governmental and societal uses of census data, is differential privacy. I have summarized here what I consider to be the most important reasons that the Census Bureau decided to adopt differential privacy.

48. **Disclosure avoidance must be proactive.** The fundamental objective of disclosure avoidance protections is to proactively prevent disclosures. Just like corporations are not expected to wait until they have suffered a major data breach before upgrading

---

DSEP is the parent organization for the Disclosure Review Board (DRB), which reviews and approves individual data releases to ensure that no confidential data is released.

[48] On May 10-11, 2017 DSEP decided that "any request for disclosure avoidance of proposed publications for the 2020 Census be routed to the 2020 DAS team before going to the DRB" meaning that all 2020 Census publications would be subject to differential privacy. See Appendices E and F. On February 15, 2018 DSEP suspended publication of "all proposed tables in Summary File 1 and Summary File 2 for the 2020 Census at the block, block-group, tract, and county level except for the PL94-171 tables, as announced in Federal Register Notice 170824806–7806–01…" acknowledging that "…these data in many cases were accurate to a level that was not supported by the actual uses of those data, and such an approach is simply untenable in a formally private system." DSEP further decided that "SF1 and SF2 will be rebuilt based on use cases." See Appendix G. In parallel with these decisions by DSEP, the disclosure risks identified by the preliminary results of the simulated reconstruction attack also led to this issue being added to the Census Bureau's risk management portfolio. On April 17, 2017 the risk of reconstruction attacks was proposed for inclusion in the Research and Methodology Directorate's risk registry. On September 12, 2017 it was escalated and included on the Enterprise-level Risk register. Finally, on January 30, 2018, it was further escalated to the Enterprise-level Issue register, with the development and use of the 2020 Census Disclosure Avoidance System as an identified resolution action to be taken. .

their IT security systems to protect against known threats, statistical agencies should not wait until they suffer a confirmed breach before improving their disclosure avoidance protections to account for known threats. The expectation, for both IT security and disclosure avoidance, is to remain vigilant about emerging threats and risks, and to take appropriate action *before* those risks lead to a breach.

49. **The privacy risk landscape has fundamentally changed since 2010.** Traditional methods of assessing disclosure risk rely on knowing what tools and resources an attacker might leverage to undermine confidentiality protections. These tools, however, are ever evolving. Over the last decade, technological advances have made powerful cloud computing environments, with sophisticated optimization algorithms capable of performing large-scale attacks, cheap and easily available. While these tools were not yet a viable attack model in 2010, they certainly represent a credible threat today.[49]

50. **Internal research has conclusively proven the fundamental vulnerabilities of the 2010 swapping methodology.** The Census Bureau has performed extensive empirical analysis of the disclosure risk inherent to the 2010 Census swapping methodology as detailed in Appendix B. No technique can produce usable data with absolutely zero risk of re-identification, but the re-identification rates from our internal experiments on the 2010 Census swapping methodology are orders of magnitude higher than what they were intended to be. The privacy threat landscape has evolved over the last decade and compels the Census Bureau to adapt its protections accordingly.

[49] DSEP drew this conclusion from the simulated reconstruction-abetted re-identification attack in Appendix B. The Office of National Statistics reached the same conclusion in its 2018 "Privacy and data confidentiality methods: a Data and Analysis Method Review (DAMR)" at Privacy and data confidentiality methods: a Data and Analysis Method Review (DAMR) – GSS (civilservice.gov.uk) (cited on April 10, 2021).

51. **The Census Bureau determined that differential privacy was the only method that could adequately protect the data while preserving the value of census data products.** When our internal research demonstrated the vulnerabilities of the swapping mechanism used for the 2010 Census, we considered a range of options for the 2020 Census. The three leading options were differential privacy, an enhanced version of data swapping, and a return to whole-table suppression. But to achieve the necessary level of privacy protection, both enhanced data swapping and suppression had severely deleterious effects on data quality and availability. With its enhanced privacy protections and precision control over the tuning of privacy/accuracy tradeoff, the Census Bureau determined that differential privacy was the only viable solution for the 2020 Census.

52. **Differential privacy can be fine-tuned to strike a balance between privacy and accuracy.** DSEP made the preliminary decision to pursue differential privacy on May 10-11, 2017. Since that decision was announced, the Census Bureau has worked extensively with our advisory committees, federal agency partners, American Indian and Alaska Native tribal leaders, the Committee on National Statistics, professional associations, data user groups, and many others at the national, state, and local levels to understand how they use decennial census data and to ensure that our implementation of differential privacy will preserve the value of the decennial census as a national resource. The Census also released sets of demonstrative data to allow the public and end-users to provide feedback that allowed us to fine-tune and tweak how we will ultimately implement differential privacy.[50]

[50] U.S. Census Bureau "Developing the DAS: Demonstration Data and Progress Metrics" https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-development.html.

53. **The need to modernize our privacy protections has been confirmed by external experts.** The Census Bureau's ongoing partnerships with scientific and academic experts from around the country helped us conduct the internal evaluation of the disclosure risk of the 2010 Census swapping methodology and confirmed the need to modernize our privacy protections. To supplement this ongoing work and to get external expert confirmation of the conclusions that we have drawn from it, the Census Bureau also commissioned an independent expert review by JASON, an independent group of elite scientists that advise the federal government on science and technology. The JASON report confirmed our findings regarding the re-identification risk inherent to the 2010 Census swapping methodology.[51]

54. **Differential Privacy can produce highly accurate data.** One key benefit of differential privacy is the ability to fine-tune privacy and accuracy. The next iteration of demonstration data will establish that differential privacy protections can produce extremely accurate redistricting data. While the full April 2021 Demonstration Data Product[52] and supporting metrics will be released by April 30, 2021, I can provide a high-level summary of key metrics:[53]

    - Total populations for counties have an average error of +/- 5 persons (reflecting a mean absolute percent error of 0.04% of the counties' population) as noise from

[51] JASON (2020). "Formal Privacy Methods for the 2020 Census" JASON Report JSR-19-2F. https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/privacy-methods-2020-census.pdf.

[52] The April 2021 demonstration data uses a global privacy-loss budget of 10.3 with a very substantial proportion allocated to detailed race and ethnicity statistics at the block and block group levels.

[53] Statistics for the April 2021 Demonstration Data Product are preliminary, based on the internal research version. The production version will be used for the detailed summary statistics when they are posted on census.gov.

differential privacy.[54] This is extremely accurate considering that if we simulate the errors in census counts as estimates of the true population, then the average county-level estimation uncertainty of the census is +/- 960 persons (averaging 1.6% of the county census counts).[55]

- At the block level the differentially private data have an average population error of +/- 3 persons, which includes both housing unit and group quarters populations. Compare that with the simulated error inherent in the census which puts the average error uncertainty of block population counts at +/- 6 people.[56]

---

[54] The statistics are the mean absolute error and the mean absolute percentage error in county population comparing the April 2021 Demonstration Data Product to the data released in the 2010 Summary File 1.

[55] The inherent error in the census counts as estimates of the true population can be simulated using data-defined person and correct-enumeration rates from coverage measurement estimates, in this case from the most recent decennial census in 2010. (See Mule, T. "2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States", Report G-10, g01.pdf (census.gov). Table 3, in particular.) An alternative modeling perspective simulates the natural variation of census population estimates using the natural variation in census estimates due to erroneous enumerations and other sources of error inherent in the Census. For county populations this natural variation is about +/- 120 persons (0.3% of population), also based on coverage data from the 2010 Census. As with all simulation estimates, there is sensitivity to the assumptions. The reported statistics are the mean absolute error and the mean absolute percentage error. Differentially private statistics include both the housing unit and group quarters populations. Simulations exclude the group quarters population because there are no coverage estimates for that group.

[56] The simulation of the natural variation of census block-level populations is +/- 1.5 persons, which excludes the group quarters population because there are no coverage estimates for that group. As with all simulation estimates, there is sensitivity to the assumptions. The reported statistics are the mean absolute errors. Mean absolute percentage errors are not useful statistics for block populations because more than 2,000,000 blocks with positive housing units have populations between 0 and 9. Differentially private statistics include both the housing unit and group quarters populations. Simulations exclude the group quarters population because there are no coverage estimates for that group.

55. **The April 2021 demonstration data show no meaningful bias in the statistics for racial and ethnic minorities** even in very small population geographies like Federal American Indian Reservations. The data permit assessment of the largest OMB-designated race and ethnicity group in each geography—the classification used by the Department of Justice for Voting Rights Act scrutiny—with a precision of 99.5% confidence in variations of +/- 5 percentage points for off-spine geographies as small as 500 persons, approximately the minimum voting district size in the redistricting plans that the Department of Justice provided as examples.
56. **The accuracy of differential privacy increases at higher levels of geography, even for arbitrary geographic areas like Congressional and legislative districts.** The Census Bureau designed its implementation of differential privacy to increase accuracy as blocks are aggregated into larger geographic areas like neighborhoods, voting districts, towns, and other places. Rather than infusing noise at the block level and aggregating upwards, which would cause error to compound at larger geographic levels, the Disclosure Avoidance System's TopDown Algorithm (TDA) takes the opposite approach. Starting at the national level, the algorithm establishes very precise (but still privacy-protected) tabulations for all characteristics at the national level, then works its way down the geographic hierarchy, ensuring that all of the geographic entities at each level (e.g., the Census tracts within a county) add up precisely to the established characteristics of the level above (e.g., the county). This approach limits the distortions that can arise from noise infusion and ensures the reliability of statistics as the underlying size of the population increases. Plaintiffs argue that "the November 2020 demonstration data also skewed the 2010 tabulations enough to create a population deviation in Alabama's Congressional districts on a level that courts have found in other contexts to violate voters' equal population rights," with districts losing up to 73 individuals or gaining 206 individuals over reported values. While this may have been true for the November 2020 Demonstration Data Product, this is not

true for the Demonstration Data Product that will be produced by the end of April. In the April 2021 Demonstration Data Product, Congressional districts as drawn in 2010 have a mean absolute percentage error of 0.06%. If the Congressional districts had been drawn using the April 2021 Demonstration Data Product, their statistical composition for the purposes of Voting Rights Act scrutiny would not be affected. Even for state legislative districts, which had average sizes of 159,000 (upper chambers) and 64,000 (lower chamber), the mean absolute percentage errors are 0.09% (upper chambers) and 0.16% (lower chambers), respectively. Such errors are trivial and imply that the difference between districts drawn from the April 2021 Demonstration Data Product and those drawn from the original 2010 P.L. 94-171 Redistricting Data Summary File would be statistically and practically imperceptible. *Most importantly for the redistricting use case, the TDA, when properly tuned, ensures that redistricters can remain confident in the accuracy of the population counts and demographic characteristics of the voting districts they draw, despite the noise in the individual building blocks.*

IMPLEMENTING DIFFERENTIAL PRIVACY FOR THE 2020 CENSUS

57. Census announced that it planned to use Differential Privacy for the 2020 Census in a few different venues: (1) August 3, 2018, 2020 Census Program Management Review; (2) December 6, 2018, Census Scientific Advisory Committee Meeting; and (3) May 2, 2019, Census National Advisory Committee meeting.

58. The Bureau has engaged in a years-long campaign to educate the user community and solicit their views about how differential privacy should be implemented. Census Bureau staff have made hundreds of public presentations, held dozens of webinars, held formal consultations with American Indian and Alaska Native tribal leaders, created an extensive website with plain English blog posts, and conducted regular outreach with dozens of stakeholder groups. We have made presentations to our

scientific advisory committees and provided substantial information to oversight entities such as the Government Accountability Office and the Office of the Inspector General.

59. Part of the Bureau's effort to inform the public and solicit feedback involved releasing a series of Demonstration Data Products. There are many different ways to implement differentially private disclosure avoidance mechanisms, and the design and parameters of these mechanisms can substantially impact the fitness-for-use of the resulting data. The Census Bureau's TopDown Algorithm (TDA) was specifically designed to address the reconstruction-abetted re-identification vulnerability risks, while allowing the Bureau to tune the accuracy of the statistics to ensure fitness-for-use.

60. To date, the Census Bureau has released four sets of Demonstration Data Products (in October 2019, May 2020, September 2020, and November 2020). The Census Bureau has received substantial, actionable feedback after each release that has contributed to the system's design and optimization.

61. All four of these demonstration products used a lower privacy-loss budget than we anticipate using for the final 2020 Census data—that is, these demonstration data were purposefully "tuned" to privacy and not "tuned" for producing highly accurate redistricting data. We held the privacy-loss budget roughly the same across these four releases to allow us to compare effects of incremental improvements in the system. After each release, these demonstration files enabled data users to help the Census Bureau identify areas where the algorithm needed to be tuned to meet their specific use cases. While the Census Bureau has not yet set the final privacy-loss budget, we have been clear that all the demonstration data released to date have used a lower

privacy-loss budget (more privacy, less accuracy) than will be selected for the final production run of the redistricting data.[57]

62. This degree of transparency into the design and implementation of a disclosure avoidance methodology is unprecedented in the federal government. The Census Bureau has submitted its differential privacy mechanisms, programming code, and system architecture to thorough outside peer review. We have also committed to publicly releasing the entire production code base and full suite of implementation settings and parameters. Many traditional disclosure avoidance methods, most notably swapping techniques, must be implemented in a "black box." Implementation parameters for these legacy disclosure avoidance methods, especially swapping rates, are often some of the most tightly guarded secrets that the Census Bureau protects. But differential privacy does not rely on the obfuscation of its implementation as a means of protecting the data. The Census Bureau's transparency will allow any interested party to review exactly how the algorithm was applied to the 2020 Census data, and to independently verify that there was no improper or partisan manipulation of the data.

INVARIANTS ARE NOT REQUIRED FOR ACCURACY.

63. Invariants—or data held constant when applying differential privacy—introduce privacy risks and are not necessary to ensure accuracy. Invariants were not well understood either theoretically or empirically in 2016 when the Census Bureau began its research on differential privacy for decennial census data, but we now understand that invariants defeat the privacy protections and must be limited in order to protect the integrity of the system as a whole. Unlike traditional approaches to disclosure avoidance, differentially private noise infusion offers quantifiable and provable privacy guarantees. These guarantees, reflected in the global privacy-loss budget and its

[57] Most recently on February 23, 2021 in The Road Ahead: Upcoming Disclosure Avoidance System Milestones (govdelivery.com).

allocation to each statistic, serve as a promise to data subjects that there is an inviolable upper bound to the risk that an attacker can learn or infer something about those data subjects through publicly released data products. While that upper bound is ultimately a policy decision, and may be low or high depending on the balancing of the countervailing obligations to produce accurate data and to protect respondent confidentiality, the level of the global privacy-loss budget is central to the ability of the approach to protect the data. Invariants are, by their very nature, the equivalent of assigning infinite privacy-loss budget to particular statistics, which fundamentally violates the central promise of differentially private solutions to controlling disclosure risk. By excluding the accuracy of invariant data elements from the control of the privacy-loss budget, invariants exclude the disclosure risk and potential inferences that can be drawn from those data elements from the formal privacy guarantees. Thus, instead of being able to promise data subjects that the publication of data products will limit an attacker to being able to infer, at most, a certain amount about them (with that amount being determined by the size of the privacy-loss budget and its allocation to each characteristic), the inclusion of one or more invariants fundamentally excludes attacker inferences about the invariant characteristic(s) from the very nature of that promise. The qualifications and exclusions to the privacy guarantee weaken the strength of the approach and make communicating the resulting level of protection substantially more difficult. This is the reason that DSEP removed the block-level invariant on population and voting-age population. Below the state level, DSEP only authorized block-level invariants that were necessary to conduct the field operations of the 2020 Census: housing unit address counts, and occupied group quarters address counts and types. As noted above, if the block population is reported with some random fluctuation around the confidential value, then only by chance will the block identifier be correct in any potential reconstructed microdata. Compound this effect

over 8,000,000 blocks and the number of feasible reconstructions explodes exponentially. This is what provides the protection against re-identification from the reconstructed data.

64. Invariants are not required to improve the accuracy of any statistic processed by differential privacy. Assigning sufficiently high (but not infinite) privacy-loss budget to any statistic can ensure perfect accuracy for that statistic while still allowing the resulting privacy-loss to be communicated in the privacy guarantee. For example, the state-level population of the American Indian and Alaska Native tribal areas has been given sufficient privacy-loss budget to ensure that those populations are presented accurate to the number of persons in the units column; the mean absolute error is 1 person, essentially invariant and the same precision as the state populations themselves. But this solution still requires balancing accuracy and privacy-loss overall. All characteristics cannot have large privacy-loss budget allocations at every geographic level. If they did, the published tables would be exact images of the confidential data and subject to the same vulnerability as the 2010 Census.
65. The forthcoming April 2021 Demonstration Data Product illustrates this tradeoff. These new demonstration data use a global privacy-loss budget for persons of 10.3, which is much larger than the 4.0 budget used in the earlier releases but is still allocated in a manner that provides a level of protection for every census record and every published characteristic. The April 2021 demonstration data also fully satisfy a tightly specified set of accuracy criteria specialized to the redisticting use case. Specifically, populations, voting-age populations, and the proportion of the largest OMB-designated race and ethnicity groups are all reliable for redistricting and Voting Rights Act scrutiny in arbitrary contiguous block aggregates for both on-spine and off-spine political and legal entities. Because new districts cannot be drawn before the 2020 P.L. 94-171 Redistricting Data Summary File is released, counties, block

groups, minor civil divisions, incorporated places, and Census-designated places were all used as on- and off-spine geographic entities for tuning purposes.

66. In the April 2021 Demonstration Data Product, all the targeted small population statistics for race and ethnic groups are far more accurate than in previous demonstration data products, even though no additional invariants were used. The gain in accuracy is entirely due to dedicating more of the privacy-loss budget to the block- and block group-level statistical tables and carefully specifying the differentially private measurements to target the OMB-designated race and ethnicity groups. Biases in the tribal areas' race and ethnicity data were also greatly reduced.

67. The Census Bureau has received substantial feedback from our data user community highlighting distortions that were present in the early versions of our demonstration data, particularly in the version released in October 2019. Based on that feedback, the Census Bureau has identified and corrected the algorithmic sources of those distortions. As these measures of accuracy and bias show, any residual impact of the types of systematic bias observed in the early demonstration data will be negligible and well within the normal variance and total error typical for a census.

PROCESS AND TIMELINE MOVING FORWARD

68. The operational delays caused by the global COVID-19 pandemic, and the resulting processing schedule changes for production of the redistricting data product shifted the milestone dates for all the systems necessary to produce the data. While the 2020 Census Disclosure Avoidance System is fully operational, and has already passed the Test Readiness Review (TRR) and Production Readiness Review (PRR) milestones on schedule, we have taken advantage of the additional time before the May 20, 2021 Operational Readiness Review (ORR) to perform additional optimization and testing of the system, and to engage in another round of data user evaluation and feedback.

69. The Census Bureau will release another demonstration product by April 30, 2021 using a higher privacy-loss budget (more accuracy) that better approximates the final privacy-loss budget that will likely be selected for the redistricting data product. These new demonstration data will also reflect system design changes that have been made since the last demonstration data release, along with tuning and optimization of the system that have been done specifically to prioritize population count accuracy and the ability to identify majority-minority districts.[58] The new release will give users yet another opportunity to let the Census know specifically where the data are (or are not yet) sufficiently accurate to meet their requirements.
70. On March 25, 2021, DSEP approved the privacy-loss budget to be used for the next demonstration product. This privacy-loss budget reflects empirical analysis of over 600 full-scale runs of the Disclosure Avoidance System using 2010 Census data. The Census evaluated these experimental runs using accuracy and fitness-for-use criteria for the redistricting use case informed by the extensive feedback we have received from the redistricting community and the Civil Rights Division at the U.S. Department of Justice. Based on this feedback, the privacy-loss budget for the final demonstration product is set to ensure the accuracy of racial demographics for voting districts as small as 500 individuals. With this tuning, the proportion of the largest racial group within even those small state/local voting districts of 500 individuals will be accurate to within five percentage points of the enumerated value at least 95% of the time. As voting district population size increases to any sort of reasonably anticipated legislative district, the error will be miniscule. For example, Congressional and

[58] Users will be able to see the difference between algorithmic improvements and greater privacy-loss budget. At the same time as the main April 2021 Demonstration Data Product is released, the Census Bureau will also release demonstration data using exactly the same software implementation but setting the global privacy-loss budget to 4.0 for persons, as it was in the four previous demonstration data products.

state legislature districts will have significantly higher accuracy for population counts and voting age population counts.

71. Following the release of the new demonstration data, data users and stakeholders will have about a month to submit additional feedback on their analysis and assessment of these data, before DSEP, in early June 2021, sets the privacy-loss budget and system parameters for the production run of the redistricting data product.

72. The production run for creating the Microdata Detail File (the internal name for the file that contains the privacy-protected data) is scheduled to occur between June 26 and July 18, 2021. This roughly three-week period is similar to the period required to implement disclosure avoidance in prior censuses and is not the cause of the delay in the delivery of the redistricting data.

73. As discussed in more detail below, any court-ordered change in the Census Bureau's implementation of disclosure avoidance would add significant time to this schedule.

BRYAN AND BARBER DECLARATIONS

74. Although I cannot set out all my observations and disagreements with the declarations of Dr. Michael Barber and Mr. Thomas Bryan in this declaration, I want to identify some key areas of dispute.

75. Dr. Barber's expert report does not adequately account for the fact that the Census Bureau's demonstration data products had a privacy-loss budget significantly lower than the expected budget that will be set for the 2020 Census. As I explained above, we purposefully set the budget lower than ones most likely to be finally chosen (set to favor privacy over accuracy), so that we could isolate the distortions and demonstrate the effectiveness of various methodological modifications. One cannot draw conclusions about the accuracy of the data the Census Bureau will release for the 2020 Census based on these demonstration products.

76. Dr. Barber is premature in drawing conclusions about the accuracy of the 2020 redistricting data before the Census Bureau has set a final privacy-loss budget, and he is further incorrect in opining on the accuracy of differential privacy without considering the relative error of alternatives. Dr. Barber focuses most of his report on the possible quality concerns of differentially private 2020 Census data releases with no attention to (1) the demonstrated privacy risks of a 2020 Census protected by legacy methods and (2) the accuracy of alternatives to differential privacy including enhanced swapping or suppression. As I show in this declaration, all disclosure avoidance systems trade-off accuracy for confidentiality protection. They must be compared to each other. Releasing the redistricting data without disclosure avoidance procedures—tabulating the Census Edited File directly—is not an option and was not done for the 1990, 2000, or 2010 Censuses.

77. Dr. Barber relies on external studies that draw incorrect conclusions and use early demonstration data products. In his declaration, Dr. Barber quotes Santos-Lozada, et al. (2020) on page 14 by saying that "[i]nfusing noise in the data, in comparison to the current disclosure avoidance system, will produce inaccurate patterns of demographic change with higher levels of error found in the calculations for non-Hispanic blacks and Hispanics. At the same time, these counts are bound to impact post-2020 districting for both federal and state elections, as well as evaluations of that redistricting. . . .[T]hese changes in population counts will affect understandings of health disparities in the nation, leading to overestimates of population-level health metrics of minority populations in smaller areas and underestimates of mortality levels in more populated ones." The Santos-Lozada et al. paper uses the October 2019 Demonstration Data Product. Therefore, its conclusions are only applicable to the state of the algorithms and the overall privacy-loss budget used for that early release. Those were neither the final algorithms nor the final privacy-loss budget. I informed the editors of the Proceedings of the National Academy of Sciences of these defects during the

peer-review process. I strongly recommended that the word "will" in the title be changed to "may" for these reasons. There is nothing statistically incorrect in the paper except for the general failure of these demographers to account for estimation error due to disclosure avoidance when doing their statistical analyses as I have noted in my own scholarly work[59] and other statisticians and computer scientists have also noted.[60] The fatal error in the Santos-Lozada et al. paper is drawing conclusions from preliminary data generated by an obsolete version of the 2020 Census DAS using obsolete settings for the privacy-loss budget and its allocation. Those conclusions are wrong and so, by extension, are those of Dr. Barber.

78. Dr. Barber's conclusions do not take into account that if the Census Bureau were forced to hold the number of people in housing units invariant at the block level, that would, in turn, require adding more noise and error to the demographic characteristics of those individuals in an effort to offset what amounts to assigning block-level populations an infinite privacy-loss budget. As I show in my declaration, doing so is unnecessary and harmful to both accuracy and confidentiality protection. The correct procedure is to set accuracy targets for meaningful aggregations then tune the disclosure avoidance procedures to meet them. This procedure is transparent when using differential privacy, but it was also done for the 2010 swapping system albeit in memos that are also protected by 13 U.S. Code §§ 8(b) & 9.

79. Furthermore, Dr. Barber's work draws incorrect conclusions about biases in rural areas and for specific small populations. In his declaration, Dr. Barber states on page 13

---

[59] Abowd, John M. and Ian Schmutte "Economic Analysis and Statistical Disclosure Limitation" *Brookings Panel on Economic Activity* (Spring 2015): 221-267. [download article and discussion, open access] [download preprint].

[60] Wasserman L. and S. Zhou "A Statistical Framework for Differential Privacy," *Journal of the American Statistical Association*, Vol. 105, No. 489 (2010):375-389, DOI: 10.1198/jasa.2009.tm08651.

that "[p]laces with fewer people (rural locations) and areas with smaller, distinctive populations (minority communities) are more likely to be impacted since these are the places where identification is more concerning, and the application of statistical noise is more likely to have a larger impact on the summary statistics derived from the altered data." He concludes on pages 13 and 14 that "...the process of differential privacy is not applied equally across the entire population. Places with fewer people (rural locations) and areas with smaller, distinctive populations (minority communities) are more likely to be impacted since these are the places where identification is more concerning, and the application of statistical noise is more likely to have a larger impact on the summary statistics derived from the altered data." This conclusion is incorrect. His analysis should say that the privacy-loss of the respondents in these small areas is being treated equally and identically to the privacy-loss of the respondents in large population areas; that is, every single respondent gets the full privacy protection afforded by the DAS—unlike the 2010 system, which only tried to protect certain households. To properly compare urban/rural statistics before and after the application of disclosure avoidance, regardless of the system, the full algorithm assigning rural/urban status must be used on both the privacy-protected and confidential data. Dr. Barber has not done this.

80. Dr. Barber's work makes incorrect assertions about the non-negativity constraint. In his declaration, Dr. Barber cites Riper, Kugler, and Ruggles (2020) on page 13 stating that "[t]he non-negativity constraint requires that every cell in the final detailed histogram be non-negative. As described above, many of the cells in the noisy household histograms will be negative, especially for geographic units with smaller numbers of households. Returning these cells to zero effectively adds households to these small places, resulting in positive bias." This point is not an accurate description of how non-negativity is being handled in the post-processing of the noisy histogram. The analysis should say that negative values are not simply being returned to zero, but

that all blocks with housing units are used to estimate the population counts subject to a non-negativity constraint on the solutions. That is, negative values are not "[r]eturning to zero," the entire 2,016 element matrix (for the redistricting data) is smoothed to a consistent, non-negative matrix for each of the 8,000,000 blocks, 275,000 block groups, 75,000 tracts, 3,143 counties, 51 states (including DC), and the U.S. simultaneously.[61] At the block-level, there are expected to be an average of only 40 people represented across the 2,016 cells. This is the inherent sparsity that any disclosure avoidance system must address. Dr. Barber claims on page 13 that "[t]he combination of the non-negativity constraint and population invariants consistently leads to bias increasing counts of small subgroups and small geographic units and decreasing counts of larger subgroups and geographic units." While the statement is correct in principle, the magnitudes shown in his report are not representative of the final redistricting data product. At the levels of privacy-loss budget used for the forthcoming April 2021 Demonstration Data Product, the consequences of the non-negativity constraint were tightly controlled for population areas of at least 500 total persons. The remaining variation in block-level statistics, including small biases, is required to protect locational privacy and deliver consistent data. It is well within the inherent variability of block-level census data, as shown in my declaration.

81. Dr. Barber argues that the amount of error observed in the demonstration files indicates that differential privacy cannot produce data sufficient for important use cases. Mr. Barber's focus on the percentage of blocks in the demonstration data that differ at all from the official 2010 Census data (even if that difference represents the addition or subtraction of a single individual from the block) ignores two important points.

[61] The matrix is 2,016 elements rather than 252 because there are eight elements in the Group Quarters Table P5 (seven group quarter types and "not a group quarters") that also interact with the other categories. The number of geographic entities at each level is based on approximate values for 2020 tabulation geographies.

First, the entire objective of our implementation of differential privacy is to infuse sufficient noise in block-level data to protect against reconstruction-abetted re-identification attacks while ensuring that when those blocks are aggregated into larger geographies of interest (voting districts, towns, etc.) those relative errors diminish and the accuracy of the tabulations improves. Second, the overall accuracy of the data is a direct consequence of the global privacy-loss budget selected and how it is allocated. The demonstration data used by both Dr. Barber and Mr. Bryan for their analyses, which use a substantially lower privacy-loss budgets than will be used for the final 2020 Census data products, can therefore be expected to be substantially "noisier" than the final data will be. Examples of noise levels in the April 2021 Demonstration Data Product provided in my report and verifiable when those data are released later this month confirm my claims.

82. Mr. Bryan assesses the accuracy of the four Demonstration Data Products (October 2019, May 2020, September 2020 and November 2020) using the percent of blocks with any change at all (pp. 9-13) or percentage errors (pp. 16-19). Both sets of analyses are based on obsolete versions of the DAS, but they also make serious errors that will still be salient when he uses the April 2021 Demonstration Data Product. The DAS was designed to control the error in counts, not percentages. The basic tables in the P.L. 94-171 Redistricting Data Summary File are counts of resident persons living in specific geographies who have features chosen from the following taxonomy {any age, voting age}, {Hispanic/Latino, not Hispanic/Latino}, and any combination of {Afro-American/Black, American Indian/Alaska Native, Asian, Native Hawai'ian/Pacific Islander, White, Some other race} except "none." The specific aggregate geographies available in the data product are all built from census blocks, but it is the counts of persons in those aggregate geographies, including voting districts, not the block counts themselves that must be accurate enough to be fit for redistricting. Block-level errors, whether in counts or percentages, are irrelevant except to the extent that they

are not controlled in larger-population geographies. In 2010, the average population in a block was 28 and the average population in an occupied block was 49. Any block-level variation in one of the 2,016 cells of the redistricting data for total populations this small is going to appear as a "large" percentage error. Indeed, most of those statistics have a base of zero, making percentage variation undefined and meaningless. The DAS must introduce noise into the block-level data to achieve any confidentiality protection at all. This statement is also true for the systems that were used in the 1970 to 2010 Census. The noise from suppression (1970, 1980) is counts that are simply not reported at the block level. The noise from blank and impute (1990) is due to the imputation modeling. The noise from swapping (2000, 2010) is due the exchange of geographic identifiers across blocks. All confidentiality protection applied to block-level redistricting data produces errors of the sort described by Mr. Bryan. Furthermore, many of the supposed DAS errors in Mr. Bryan's analysis cancel out when blocks are aggregated into larger-population geographies like block groups, census tracts, towns, counties, and congressional districts. This is not an accident; it is a carefully designed feature of the DAS. The tabulation of the protected microdata might miss a person in one block, but have an "excess" person in the neighboring block for a particular characteristic. Because the DAS uses direct measurements from the U.S. all the way down to the block to estimate the counts at every level of geography, whether on- or off-spine, they are all much more accurate than any of the block estimates that comprise them. This is easy to see in any balanced summary of the accuracy of the DAS. Counties and places have far smaller percentage errors than the average percentage error of the blocks that compose them.

CLARIFYING STATEMENT QUOTED IN COMPLAINT

83. Plaintiffs assert, quoting an article in 2018 by the demographer Steven Ruggles and others, that I claimed that database reconstruction does not pose a significant re-identification threat. I made the statement that plaintiffs reference indirectly at the December 14, 2018 meeting of the Federal Economic Statistics Advisory Committee (FESAC) in my own presentation.[62] Dr. Ruggles was on the FESAC program in the same session. I made the remarks in December 2018 as a report on ongoing research.[63] At the February 16, 2019 session of the American Association for the Advancement of Science (AAAS), I retracted my tentative conclusion about re-identification based on additional research reported there. The full text and presentation of the AAAS session are attached as Appendices H and I.[64] To be clear, the Census Bureau's simulated reconstruction attack on the 2010 Census data described in this declaration and in the accompanying appendix materials shows there is a significant re-identification risk. However, the Census Bureau's Data Stewardship Executive Policy Committee (DSEP) acted to adopt differential privacy as soon as that research showed that an accurate microdata reconstruction was feasible. It did not require, nor should it have required, the subsequent demonstration that those reconstructed microdata permit between 52 and 179 million correct re-identifications from the 2010 Census. The reconstructed mi-

---

[62] Federal Economic Statistics Advisory Committee program: FESAC Meeting Agenda December 2018 (bea.gov).

[63] My remarks at the December 18, 2018 FESAC: Microsoft PowerPoint - Abowd Presentation (bea.gov).

[64] AAAS materials for the February 16, 2019 session area also here: https://blogs.cornell.edu/abowd/files/2019/04/2019-02-16-Abowd-AAAS-Talk-Saturday-330-500-session-FINAL-as-delivered-2jr4lzb.pdf and https://blogs.cornell.edu/abowd/files/2019/04/2019-02-16-Abowd-AAAS-Slides-Saturday-330-500-session-FINAL-as-delivered-1iqsdg2.pdf.

crodata fail the *2010 Census* microdata disclosure avoidance requirements—the requirements that were in place for that census—because they contain geographic identifiers (the block code) that relate to a minimum population of one rather than the 100,000 person minimum population that contemporary standards required. The reconstructed microdata also did not impose any of the minimum population thresholds required of the tabulation variables, especially age.[65] These requirements were already in place because it is well understood at the Census Bureau and in the official statistics community worldwide that geographic identifiers for low-population areas, sex, and exact age in microdata files are a major disclosure risk especially in population censuses.

IMPACT OF ANY COURT RULING BARRING USE OF DIFFERENTIAL PRIVACY

84. Were the Court to rule that the Census Bureau was precluded from using differential privacy for the 2020 Census P.L. 94-171 Redistricting Data Summary File, we would be faced with hard choices. The inevitable result would be significant delay in delivery of the already-delayed redistricting data and diminished accuracy. Either the Census Bureau would have to revert to using suppression (as was last used in the 1980 Census) or use enhanced swapping (as was used in the 1990 to 2010 Censuses, but at a much higher rate and with fewer invariants). Either choice would delay results and diminish accuracy.

85. The effect on the schedule for delivering redistricting data would be substantial. The Census Bureau cannot ascertain the length of the delay until it understands any parameters the Court might place on its choice of methodology, but under all scenarios the delay would be multiple months. This delay is unavoidable because the Census Bureau would need to develop and test new systems and software, then use them in

[65] McKenna (2019a).

production and subject the results to expert subject matter review prior to production of data. The Census Bureau has been developing the systems and software to use differential privacy for several years—the agency has spent millions of dollars purchasing cloud computer capacity and writing and tuning code. The systems and software are ready to go and await only final tuning and a decision on the privacy-loss budget.

86. Even if the agency was ordered to repeat exactly what was done in 2010 (despite the serious risks to privacy the Census has identified), we could not simply "flip a switch" and revert to the prior methodology. Instead, we would need to conduct the requisite software development and testing. The 2020 Census's system architecture is completely different than that used in the 2010 Census, and it is thus not possible to simply "plug in" the disclosure-avoidance system used in 2010.

87. Not only would redistricting data be further delayed, but the resulting data would be less accurate. Both swapping and suppression are blunt instruments for privacy protection. Unlike differential privacy, neither can be effectively tuned to optimize for data accuracy. Knowing that the 2010 Census results were vulnerable to reconstruction, the Census Bureau cannot simply repeat the swapping protocols from the 2010 census, but rather would be forced to fashion appropriate levels of protection for either system. Using an appropriate level of protection for either suppression or swapping would produce far less accurate data than would differential privacy.

88. I would urge any court to be quite wary of opining on the suitability of particular methods for conducting disclosure avoidance, as these decisions are highly technical and can have unanticipated consequences. The only reason the Court knows so much about the proposed methods for the 2020 Census is that transparency does not undermine their confidentiality protections, which is not the case for either swapping or suppression. While we cannot predict the full impact of any change, there is a danger than any change would have cascading effects on data accuracy and privacy, making

race and ethnicity data, along with age data, substantially less accurate. Any sort of change in the basic methodology would be minimally tested and would not have the benefit of any input from the user community.

89. In conclusion, it is my professional opinion that the Census Bureau's Data Stewardship Executive Policy Committee should be permitted to control the type and parameters of any disclosure avoidance system used for the 2020 Census, just as it did for the 2010 Census and just as its predecessor committees did for decennial censuses conducted since the passage of the Census Act (13 U.S. Code) in 1954.

I declare under penalty of perjury that the foregoing is true and correct.

DATED and SIGNED:

JOHN ABOWD Digitally signed by JOHN ABOWD
Date: 2021.04.13 10:18:58 -04'00'

John M. Abowd

Chief Scientist and Associate Director for Research and Methodology

United States Bureau of the Census