Differential Privacy and Census Data:
Implications for Social and Economic Research[†]

By Steven Ruggles, Catherine Fitch, Diana Magnuson,
and Jonathan Schroeder*

In September 2018, the Census Bureau announced a new set of methods for disclosure control in public use data products, including aggregate-level tabular data and microdata derived from the decennial census and the American Community Survey (ACS) (US Census Bureau 2018a). The new approach, known as differential privacy, "marks a sea change for the way that official statistics are produced and published" (Garfinkel, Abowd, and Powazek 2018).

In accordance with census law, for the past six decades the Census Bureau has ensured that no census publications allow specific census responses to be linked to specific people. Differential privacy requires protections that go well beyond this standard; under the new approach, responses of individuals cannot be divulged even if the identity of those individuals is unknown and cannot be determined. In its pure form, differential privacy techniques could make the release of scientifically useful microdata impossible and severely limit the utility of tabular small-area data.

Initially, the Census Bureau plans to apply differential privacy techniques to the two most intensively-used sources in social science and policy research, the ACS and the decennial census (US Census Bureau 2018b). These data generate some 17,000 publications each year. The ACS and decennial census are widely used in analyses of the economy, population change, and public health, and they are indispensable tools for federal, state and local planning. Common topics of analysis include poverty, inequality, immigration, internal migration, ethnicity, residential segregation, transportation, fertility, nuptiality, occupational structure, education, and family change. The data are routinely used to construct contextual measures that control for neighborhood effects on health and disease. Investigators exploit policy discontinuities across time and space, disasters, and weather events as natural experiments that allow causal inferences. Policymakers and planners use small-area data from the ACS and decennial census to understand local environments and focus resources where they are needed. Businesses use the data to estimate future demand and determine business locations.

Adoption of differential privacy will have far-reaching consequences for users of the ACS and decennial census. It is possible—even likely—that scientists, planners, and the public will soon lose the free access we have enjoyed for the past six decades to reliable public Census Bureau data describing American social and economic change.

The differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau (Ruggles et al. 2018). By imposing unrealistic disclosure rules, the Census Bureau may be forced to lock up data that are indispensable for basic

research and policy analysis. If public use data become unusable or inaccessible because of overzealous disclosure control, there will be a precipitous decline in the quantity and quality of evidence-based policy research.

I. Differential Privacy and Census Law

Differential privacy guarantees that the presence or absence of any individual case from a database will not significantly affect any database query. In particular, "even if the participant removed her data from the dataset, no outputs … would become significantly more or less likely" (Dwork 2006, p. 9). This definition has the advantage of being relatively simple to formalize, and that formalization yields a metric summarizing a database's level of "privacy" in a single number (ϵ).

The application of differential privacy to census data represents a radical departure from established Census Bureau confidentiality laws and precedents (Ruggles et al. 2018). The differential privacy requirement that database outputs do not significantly change when any individual's data is added or removed has profound implications. In particular, under differential privacy it is prohibited to reveal characteristics of an individual even if the identity of that individual is effectively concealed.

As the Census Bureau acknowledges, masking respondent characteristics is not required under census law. Instead, the laws require that the identity of particular respondents shall not be disclosed. In 2002, Congress explicitly defined the concept of identifiable data: it is prohibited to publish "any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means."[1]

For the past six decades the Census Bureau disclosure control strategy has focused on targeted strategies to prevent re-identification attacks, so that an outside adversary cannot positively identify which person provided a particular response. The protections in place—sampling, swapping, suppression of geographic information and extreme values, imputation, and perturbation—have worked extremely well to meet this standard. Indeed, there is not a single

---

[1] Title 5 USC. §502 (4), Public Law 107–347.

documented case of anyone outside the Census Bureau revealing the responses of a particular identified person in public use decennial census or ACS data.

II. Reconstruction and Re-identification

Census analysts argue that new disclosure rules are needed because of the threat of "database reconstruction." Database reconstruction is a process for inferring individual-level responses from tabular data. Abowd (2017) argues that database reconstruction "is the death knell for public-use detailed tabulations and microdatasets as they have been traditionally prepared."

The Census Bureau conducted a database reconstruction experiment that sought to identify the age, sex, race, and Hispanic origin for the population of each of the 6.3 million inhabited census blocks in the 2010 census. According to Abowd (2018a), the experiment confirmed "that the micro-data from the confidential 2010 Hundred-percent Detail File (HDF) can be accurately reconstructed" using only the public use summary tabulations. The HDF is the individual-level complete census incorporating confidentiality protections such as swapping similar households that reside in different places.

It should not be a great surprise that individual-level characteristics can be inferred from tabular data. Any table that includes data about people can be rearranged as individual-level data. For the Census Bureau database reconstruction experiment, analysts started with a table of age by sex by race by Hispanic origin, and converted the table to microdata. For example, if a particular census tract had three black non-Hispanic women aged 25 to 29, they created three microdata records with these individual-level characteristics. By repeating this process for every cell in the table, the full content of the table may be expressed in the form of microdata. Then the Census Bureau added more detail on place of residence, age, and race by cross-referencing across multiple tables.

The reliability of the method varies depending on the characteristics of the census block. For some blocks, there are multiple possible solutions, making inferences difficult (Abowd 2018b). In other cases it is easy to infer

individual-level variables. For example, 47 percent of blocks contain a single race and 60 percent have a single Hispanic (or non-Hispanic) ethnicity; accurately inferring race or ethnicity for persons in such homogeneous blocks is trivial. Once the individual-level data were fully reconstructed, the Census Bureau tested the accuracy by matching the reconstructed individual-level records to the microdata that had been used to create the public use tables. For each individual in the reconstructed dataset, the software searched the original microdata for a person with a matching age, sex, race, and Hispanic origin.

In the end, only 50 percent of the reconstructed cases accurately matched a case from the HDF source data (Abowd 2018c; Hansen 2018). In the great majority of the mismatched cases, the errors resulted from a discrepancy in age. Given the 50 percent error rate, it is not justifiable to describe the microdata as "accurately reconstructed" (Abowd 2018a).

Reconstructing microdata from tabular data does not itself allow identification of respondents; to determine who the individuals actually are, one would then have to match their characteristics to an external identified database (including, for example, names or Social Security numbers) in a conventional re-identification attack. The Census Bureau attempted to do this but only a small fraction of re-identifications actually turned out to be correct, and Abowd (2018d) concluded that "the risk of re-identification is small." Therefore, the system worked as designed: because of the combination of swapping, imputation and editing, reporting error in the census, error in the identified credit agency file, and errors introduced in the microdata reconstruction, there is sufficient uncertainty in the data to make positive identification by an outsider impossible.

III. Implications for Tabular Data

Despite the low risk of re-identification in the Census Bureau experiment, the 100 percent tabular data from the decennial census pose some special disclosure control challenges. Because these tables include the entire population with very fine geographic detail, there could be potential for re-identification if no disclosure protections were applied.

The block-level decennial tables include very few variables, and the research applications of these tables are comparatively limited. The Census Bureau has not yet demonstrated that differential privacy is the most effective and efficient means of preventing positive re-identification while maximizing utility of these data. It is nevertheless possible that some variant of differential privacy or a similar method could be applied that would preserve usability for the relatively limited applications of the block data while strengthening disclosure control.

Differentially-private tabular data from the ACS is considerably more challenging than the 100 percent files, because there are many more variables and the data are used for a much wider range of research and planning purposes. It may be impossible to create a differentially-private version of the ACS tables that would meet the needs of researchers and planners. Fortunately, tabular data from the ACS have features that make them inherently less identifiable than the 100 percent census data. The ACS is a sample with just 1.5 percent of the population each year, and there is no block-level data. At the block group level, the ACS data must combine five years of data, so there is temporal as well as spatial uncertainty. The chances of any particular respondent being included in the file are very low. If an exact match is found through a reconstruction and re-identification attack, it would be impossible to determine whether the match was correct because there may be another exact match which was not sampled. Accordingly, less aggressive disclosure controls may be appropriate for ACS tabular data.

IV. Implications for Microdata

Differentially private microdata is not a realistic disclosure control solution. ACS microdata samples directly provide individual-level characteristics derived from real people, and this in itself represents a violation of the core principles of differential privacy (Bambauer, Krishnamurty, and Sarathy 2014). A recent paper published by Census Bureau privacy experts notes that "record-level data are exceedingly difficult to protect in a way that offers real privacy protection while leaving the data useful for unspecified analytical purposes. At present, the Census Bureau advises research users who

require such data to consider restricted-access modalities," in particular the Federal Statistical Research Data Centers (Garfinkel, Abowd, and Powazek 2018). By "real privacy protection," the authors mean differential privacy, not confidentiality protection as defined in census law and precedent. By "unspecified analytical purposes" the authors mean any analytic purposes that are not anticipated in advance.

To guarantee differential privacy, microdata must be simulated using statistical models rather than directly derived from the responses of real people (Dajani et al. 2017, Reiter forthcoming). Such modeled data—usually called synthetic data—captures relationships between variables only if they have been intentionally included in the model. Accordingly, synthetic data are poorly suited to studying unanticipated relationships, which would greatly impede new discoveries from differentially private microdata.

Census Bureau privacy researchers argue that if the public use data become unusable, scientific research can be carried out in the secure Federal Statistical Research Data Centers (FSRDCs). This is not a practical plan. As we have argued elsewhere, the FSRDC network would have to be expanded by several orders of magnitude to accommodate the volume of research now carried out using public use microdata, and most projects would be ineligible (Ruggles et al. 2018). Without major legal changes and a massive infusion of funds, restricted access is not a viable alternative to public use microdata.

The existing ACS microdata samples provide powerful protections against re-identification. The public use microdata are a sample of a sample; annual information on less than 1 percent of the population is released to the public. There is no geographic identification of places with fewer than 100,000 inhabitants. Outlying values are top-coded or bottom-coded; variables are grouped into categories representing at least 10,000 persons in the general population; ages are perturbed for some population subgroups; and additional noise is added for persons in group quarters or with rare combinations of characteristics. These measures have proven highly effective. It is impossible for an intruder to determine whether any attempted re-identification was successful, or even to calculate the odds that the attempt was successful. Accordingly, we recommend only incremental improvements in disclosure control for the ACS microdata samples.

V. Discussion and Recommendations

There are compelling reasons to take confidentiality protection seriously. Re-identification is a greater concern today than in the past, both because of the declining cost of computing and the increasing availability of private-sector identified data that might be used in an attack. For the past two decades, the Census Bureau has conducted systematic evidence-based research on the actual risks of re-identification in public use census data (Ruggles et al. 2018). This empirical approach targets methods of disclosure control that address realistic threats by focusing on particular population subgroups and variables posing the greatest risks, while minimizing damage to data utility. The Census Bureau should build on this work by continuously modernizing and strengthening its disclosure control methods.

Differential privacy goes far beyond what is necessary to keep data safe under census law and precedent. Differential privacy focuses on concealing individual characteristics instead of respondent identities, making it a blunt and inefficient instrument for disclosure control. As Abowd and Schmutte (2019) have observed, there is a trade-off between privacy and data usability. As defined by census law, privacy means protecting the identity of respondents from disclosure. The core metric of differential privacy, however, does not measure risk of identity disclosure (McClure and Reiter 2012). Because differential privacy cannot assess disclosure risk as defined under census law and precedent, it cannot be used to optimize the privacy/usability trade-off.

The United States is facing existential challenges. We must develop policies and plans to adapt to accelerating climate change; that will require reliable ACS microdata and small area data. The impact of immigration—one of the most divisive issues in American policy debates—cannot be measured without the ACS tables and microdata. More broadly, investigators need data to investigate the causes and consequences of rapidly growing inequality in income and education. We need to examine how fault lines of race, ethnicity, and gender are dividing the country. We need basic data to study the shifts in spatial organization of the

population that are contributing to fragmentation of politics and society. This is not the time to impose arbitrary and burdensome new rules, with no basis in law or precedent, which will sharply restrict or eliminate access to the nation's core data sources.

The Census Bureau's mission is "to serve as the nation's leading provider of quality data about its people and economy" (US Census Bureau 2018c). To meet that core responsibility, the Census Bureau must make accurate and reliable data available to the public. The Census Bureau has an extraordinary record—better than anywhere else in the world—of making powerful public use data broadly accessible. Just as important, the Census Bureau also has an unblemished record of protecting confidential information. There are no documented instances in which the identity of a respondent to the decennial census or ACS has been positively identified by anyone outside the Census Bureau using public use data. We must ensure that both of these powerful traditions continue. We need both broad democratic access to high-quality data and strong confidentiality protections to understand and overcome the daunting challenges facing our nation and the world.

We have three specific recommendations:

(i) *Differential privacy might be feasible for summary files, but more testing is needed.* The most plausible use of the technique is for the 100 percent tabular files, where the range of applications is relatively limited. Making useful differentially private ACS tabular data will be challenging and may not be practical.

(ii) *To preserve the utility of public use microdata, the Census Bureau should pursue alternative disclosure control strategies.* Differential privacy is not appropriate for ACS microdata. Differentially private synthetic microdata are not suitable for most original research problems. There is no legal mandate for differential privacy, and restricted-access alternatives to public use data are not feasible.

(iii) *The Census Bureau should proceed cautiously in close consultation with the user community.* If new disclosure control technology is rushed out prematurely and without adequate evaluation, damaging mistakes are inevitable. For any new disclosure control procedures, the research community should have an opportunity to test the methods through a rigorous process before they are finalized. The best way to achieve this is by enlisting the research community to replicate past peer-reviewed research using data that incorporate new disclosure control methods.

REFERENCES

Abowd, John M. 2017. "Research Data Centers, Reproducible Science, and Confidentiality Protection: The Role of the 21st Century Statistical Agency." Paper presented at the Census Scientific Advisory Committee Meeting, Suitland, MD.

Abowd, John M. 2018a. "How Modern Disclosure Avoidance Methods Could Change the Way Statistical Agencies Operate." Paper presented at the Federal Economic Statistics Advisory Committee Meeting, Suitland, MD.

Abowd, John M. 2018b. "Staring-Down the Database Reconstruction Theorem." Paper presented at the Joint Statistical Meetings, Vancouver, BC.

Abowd, John M. 2018c. Personal communication, December 11, 2018.

Abowd, John M. 2018d. "The U.S. Census Bureau Adopts Differential Privacy." Paper presented at the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London.

Abowd, John M., and Ian M. Schmutte. 2019. "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices." *American Economic Review* 109 (1): 171–202.

Bambauer, Jane, Krishnamurty Muralidhar, and Rathindra Sarathy. 2014. "Fool's Gold: An Illustrated Critique of Differential Privacy." *Vanderbilt Journal of Entertainment and Technology Law* 16 (4): 701–55.

Dajani, Aref N., Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, et al. 2017. "The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau." Paper presented at the Census Scientific Advisory Committee Meeting, Suitland, MD.

Dwork, Cynthia. 2006. "Differential Privacy." In *Automata, Languages and Programming: 33rd International Colloquium*, edited by Michele Bugliesi, Bart Preneel, Vlarimiro Sassone, and Ingo Wegener, 1–12. Heidelberg: Springer.

Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. "Issues Encountered Deploying Differential Privacy." In *2018 Workshop on Privacy in the Electronic Society*, edited by Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter, 133–37. New York: ACM.

Hansen, Mark. 2018. "To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data." *New York Times*, December 5. https://www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html.

McClure, David, and Jerome P. Reiter. 2012. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." *Transactions on Data Privacy* 5: 535–52.

Reiter, Jerome P. Forthcoming. "Differential Privacy and Federal Data Releases." *Annual Review of Statistics and Its Application.*

Ruggles, Steven, et al. 2018. "Implications of Differential Privacy for Census Bureau Data and Scientific Research." Minnesota Population Center Working Paper 2018-6.

US Census Bureau. 2018a. "Statistical Safeguards." Data Protection and Privacy Program. https://www.census.gov/about/policies/privacy/statistical_safeguards.html.

US Census Bureau. 2018b. "Restricted-Use Microdata." US Census Bureau, https://www.census.gov/research/data/restricted_use_microdata.html#CRE1.

US Census Bureau. 2018c. *Strategic Plan—Fiscal Year 2018 through Fiscal Year 2022*. Suitland, MD: US Census Bureau.