Differential Privacy and Census Data: Implications for Social and Economic Research

Steven Ruggles†

American Economic Association
Atlanta, Georgia
January 2019

This report was prepared with the assistance of Margo J. Anderson (University of Wisconsin-Milwaukee), Jane Bambauer (Arizona State University), Michael Davern (NORC), Reynolds Farley (University of Michigan), Catherine Fitch (ISRDI), Miriam L. King (ISRDI), Diana Magnuson (Bethel University), Krish Muralidhar (University of Oklahoma), Jonathan Schroeder (ISRDI), Matthew Sobek (ISRDI), David Van Riper (ISRDI), and John Robert Warren (University of Minnesota). We are grateful for the comments and suggestions of Trent Alexander (ICPSR), Wendy Baldwin (former PRB President), John Casterline (Ohio State University), Sara Curran (University of Washington), Roald Euller (RAND Corporation), Katie Genadek (University of Colorado), Wendy Manning (Bowling Green State University), Douglas Massey (Princeton University), Robert McCaa (ISRDI), Frank McSherry, Samuel Preston (University of Pennsylvania), and Stewart Tolnay (University of Washington).

†Address correspondence to Steven Ruggles, University of Minnesota, Minnesota Population Center, 50 Willey Hall, 225 19th Ave S., Minneapolis, MN 55455 (email: ruggles@umn.edu). Support for this work was provided by the Minnesota Population Center at the University of Minnesota (P2C HD041023).

Abstract

The Census Bureau has announced a new set of standards and methods for disclosure control in public use data products. The new approach, known as differential privacy, represents a radical departure from current practice. In its pure form, differential privacy techniques may make the release of useful microdata impossible and severely limit the utility of tabular small-area data. Adoption of differential privacy will have far-reaching consequences for research. It is possible—even likely—that scientists, planners, and the public will lose the free access we have enjoyed for six decades to reliable public Census Bureau data describing American social and economic change. We believe that the differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau.

In September 2018, the Census Bureau announced a new set of standards and 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) (U.S. 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 et al. 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 policy the 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 may 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 standards to the two most intensively-used sources in social science and policy research, the ACS and the decennial census.[1] These sources account for an extraordinary volume of publications—some 17,000 each year—on the economy, population change, and public health, and they are indispensable tools for federal, state and local planning.

Adoption of differential privacy will have far-reaching consequences for users of these data. According to Census Bureau Chief Scientist John Abowd (2017), "all data publication inherently involves some inferential disclosure." Abowd maintains that this "is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared." It is possible—even likely—that scientists, planners, and the public

[1] Eventually, the Census Bureau plans to extend differential privacy to all public use data products produced by the agency (U.S. Census Bureau 2018b).

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.

We believe that the differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau. By imposing an unrealistic privacy standard, the Census Bureau may be forced to lock up data that are indispensable for basic research and policy analysis. Accurate data form the empirical foundation for social and economic models. Such data are essential for testing theories of past change, understanding present conditions, and making projections into the future.

The pages that follow review the history of census privacy policies; discuss the differences between differential privacy and census law and precedent; describe the Census Bureau's rationale for imposing the new rules; explain the special challenges of differential privacy for microdata files; explore the implications of the new approach for scientific research and planning; and discuss the important role of public perception of Census Bureau confidentiality promises. We conclude with a set of recommendations.

History of Census Privacy Policy

The 1790 Census Act specified that upon completing their enumeration of a district, each Assistant Marshall shall "cause a correct copy" of the census returns "to be set up at two of the most public places" in his district, "there to remain for the inspection of all concerned" (Wright and Hunt 1900: 14). The idea was that copies of the census returns posted in the local post office or tavern would enable members of the public to spot errors or omissions in the enumeration.

The requirement to publicly post census returns remained in effect for the next half century. The earliest concern for the confidentiality of U.S. statistical data began in 1850,

when the Secretary of the Interior declared that henceforth census returns were to be "exclusively for the use of the government, and not to be used in any way to the gratification of curiosity, the exposure of any man's business or pursuits, or for the private emolument of the marshals or assistants" (Wright and Hunt 1900: 150).

Over the course of the next century, confidentiality procedures grew increasingly rigorous. By 1880, enumerators were required to swear an oath not to disclose any information to anyone except their supervisors (Wright and Hunt 1900: 66). In 1910 President Taft made a proclamation that unequivocally promised confidentially for all census information collected (*Chicago Tribune* 1910). But the right to privacy was still not absolute; the Director of the Census had the authority to release data on individuals for "worthy purposes." During World War I, "personal information for several hundred young men was released to courts, draft boards, and the Justice Department" (Barabba 1975: 27). As late as 1921, the Director allowed a private literacy campaign to use the census to identify illiterates (General Accounting Office 1998).

In 1929 the census law authorizing the 1930 census made the confidentiality promise explicit:

> The information furnished under the provisions of this Act shall be used only for the statistical purposes for which it is supplied. No publication shall be made by the Census Office whereby the data furnished by any particular establishment or individual can be identified, nor shall the Director of the Census permit anyone other than the sworn employees to examine the individual reports (Reapportionment Act of 1929, CR 28 § 11).

With this law in place, the Census Bureau began to deny all access to data that identify particular individuals, even when the request came from another government agency. For example, in 1930 the Bureau turned down a request from the Women's Bureau for the names and addresses of employed women.

During World War II, Congress repealed the confidentiality protections in the 1929 Census Act. The Second War Powers Act (50 U.S.C. § 1402) specified that "notwithstanding any other provision of law, any record, schedule, report, or return, or any information or data contained therein, now or hereafter in the possession of the Department of Commerce, or any bureau or division thereof, may be made available by the Secretary of Commerce to any branch or agency of the Government." Under this provision, the Census Bureau made personally identified individual-level information on businesses and individuals available to other federal agencies, including surveillance agencies, from 1942 to 1947. The provision was repealed in 1947, restoring the confidentiality provisions of the 1929 statute (Anderson and Seltzer 2007, 2009).

In 1954, census law was consolidated in Title 13, which incorporated virtually the same language on confidentiality that had first appeared in the 1929 census law. In 1962 the law was strengthened to clarify that the Census Bureau was prohibited from sharing information about individuals or establishments with any other branch of government, and that census returns are "immune from legal process" and cannot be admitted as evidence in a court of law (Title 13 U.S.C. § 9, Public Law 87-813). No evidence of a breach of the Title 13 confidentiality rules during the subsequent 50 years has been uncovered (Anderson and Seltzer 2009).

Until the early 1960s, census data were disseminated exclusively through printed volumes. This format imposed practical limits on the amount of detail presented and limited threats to confidentiality. In 1962, the Census Bureau released the first electronic publication of census data, providing individual-level records (or microdata) drawn from

the 1960 “long form” census, a detailed survey filled out by one in four households. The documentation explained that the data were compliant with census privacy law:

> The one-in-a-thousand sample makes available reels of magnetic tape or sets of punchcards containing the separate records of the characteristics of a 0.1 percent sample of the population of the United States as recorded in the 1960 census. The names of the respondents and certain more detailed items on place of residence are not revealed. Therefore, it has been determined that making records available in this form does not violate the provision of confidentiality under which the census was conducted (U.S. Census Bureau 1962: 2).

The release of individual-level information was not seen as a violation of Title 13 because the Bureau did not reveal the identity of particular individuals. In addition to removing names and addresses, the Census Bureau suppressed geographic detail below the state level and top-coded income to prevent the identification of high-income persons. Only one of every 1000 persons was included in the sample, and there was no way for an outsider to determine whether any particular individual was represented.

In the mid-1960s the Census Bureau also began distributing summary tapes containing tabular data. These tapes included tables that the Census Bureau had prepared as an intermediate step in creating the 1960 census publications, providing more detail than was available in the published data (U.S. Census Bureau 1964, 1967a, 1967b). To protect confidentiality, the Census Bureau suppressed the data for geographic units with a very small population count (Courtland 1985).

The basic methods of privacy protection in Census Bureau data products remained essentially similar from 1960 to 1980 for both microdata and tabular files, although the details varied by year. Public Use Microdata Samples (PUMS) were protected by (a) stripping off names and other identifying information, (b) providing only a sample of the original data, (c) suppressing detailed geographic information, (d) top-coding continuous

variables such as income, and (e) collapsing some very detailed categories such as place of birth (Federal Committee on Statistical Methodology 2005). The tabular information was protected by suppressing the data in places with very small populations or with few members of particular subgroups.

In 1990, there was a major innovation, as described in the Census Bureau's *Monograph on Confidentiality and Privacy:*

> The data from [a sample of] households were swapped with data from other households that had identical characteristics on a certain set of key variables but were from different geographic locations. Which households were swapped was not public information. … All tables were produced from this altered file. (U.S. Census Bureau 2001a: 28).

The point of swapping is to introduce uncertainty. Swapping ensures that the information provided by a respondent cannot be confidently linked to a particular identified individual. "Because of data swapping, users should not assume that tables with cells having a value of one or two reveal information about specific individuals" (U.S. Census Bureau 2003: 8-3).

For the long-form sample questionnaire, the 1990 census employed an additional confidentiality measure: blanking and imputation. For one household in each block group, some specific values were blanked out and imputed with values "donated" from similar individuals. The imputed data were used to produce both the tabulations of long-form data and the 1990 PUMS, providing an additional layer of protection against disclosure.

Leading up to Census 2000, some Census Bureau analysts became concerned about potential disclosure risks, especially for microdata (Zayatz 1999). They argued that increasing availability of digital data—such as voter registration lists and commercial databases—together with declining costs of computing, had increased the risks of re-identification. In a re-identification attack, an external dataset that identifies particular

individuals is matched to the census microdata file. Although the use of sampling, swapping, and imputation made it impossible to identify anyone's census responses with certainty, the Census Bureau nevertheless wanted to further strengthen privacy protections. Accordingly, the Census Bureau proposed to create microdata samples with far less detail than had been available in previous census years. For example, instead of the 298 specific countries of birth identified in the 1990 census, the Bureau proposed to provide information on only the major continent of birth (Robbin 2001; Ruggles et al. 2000; Ruggles 2000).

After extensive feedback from the user community, the Census Bureau modified its plans (Robbin 2001). All variable categories representing fewer than 10,000 persons in the general population were combined into larger categories. The swapping procedure was modified to focus on cases with the highest risks of disclosure, especially persons or households that were unique within a small area. In addition, the Census Bureau used a perturbation procedure to randomly modify some ages.[2]

Similar procedures were subsequently used for the ACS, which replaced the long form of the census after 2000. Guided by on empirical re-identification experiments, the Census Bureau has continued to refine disclosure controls to further reduce the risk of re-identification. These include altering the swapping routine, better identifying households that could pose a re-identification risk, and slightly increasing the percentage of households that are swapped (Lauger et al. 2014).

[2] Unfortunately, the age perturbation was poorly executed, and the ages of persons aged 65 and older were badly skewed in both the Census 2000 microdata and the early ACS microdata (Alexander, Davern, and Stevenson 2010; Cleveland et al. 2012). The Census Bureau corrected the overall distribution in 2009 and 2010.

Differential Privacy and Census Law

As statisticians grappled with privacy in the late twentieth century, they defined a private database as one in which it was impossible to learn anything about any individual that was not already known (Dalenius 1977). As Dwork (2006) observed, it is possible for a database to reveal something about an individual even if that individual is not included in the database. For example, she argued, if we know in advance that an individual is two inches shorter than the average height, any database revealing average height would provide us with new information about that individual, even if that person was not represented in the database.

To sidestep this problem, Dwork et al. (2006) proposed differential privacy. Instead of guaranteeing that nothing about an individual can be learned from a database, differential privacy guarantees that the presence or absence of any individual case from a database should not significantly affect any database query. In particular, "even if the participant removed her data from the data set, no outputs … would become significantly more or less likely" (Dwork 2006: 9). This definition had the advantage of being relatively simple to formalize, and that formalization yielded 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 privacy laws and precedents. 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. Virtually all Census Bureau microdata and small area data products currently fail to meet that standard.[3]

As the Census Bureau acknowledges, masking respondent characteristics is not required under census law. Instead, the laws require that the identity of respondents shall not be disclosed. According to the law authorizing the census, "the Census Bureau shall not make any publication whereby the data furnished by *any particular establishment or individual* … can be identified." (Title 13 U.S.C. § 9(a)(2), Public Law 87-813). 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" (Title 5 U.S.C. §502 (4), Public Law 107–347).

The meaning of the law is clear and unambiguous: census publications must ensure that the responses of particular identified persons cannot be determined from census publications. To comply with the law, it is not necessary to mask the *characteristics* of individuals; rather, it is necessary to mask the *identity* of individuals. Thus, 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

[3] The Census Bureau plans to introduce sufficient random noise into every variable to ensure that the statistical results of a database do not change "too much" depending on whether an individual is included or excluded (Abowd and Schmutte forthcoming). The level of differential privacy in a database is defined by a single parameter ϵ that summarizes the level of noise infusion across all variables. Abowd and Schmutte (forthcoming: 5) explains that "a differentially private algorithm guarantees that the published statistics will not change 'too much' whether any observation from the confidential data is included or excluded. The notion of 'too much' is quantified by a parameter, ϵ, which measures the maximum difference in the log odds of observing any statistic across similar databases." Unfortunately, ϵ turns out to be unreliable as a measure of disclosure risk (McClure and Reiter 2012).

perturbation—have worked extremely well to meet this standard. Indeed, there is not a single documented case of anyone outside the Census Bureau revealing the responses of a particular identified person by breaking into public use decennial census or ACS data.

Database Reconstruction and Re-Identification

In September 2018, the Census Bureau announced the biggest revision of disclosure rules in at least 50 years. Census analysts have explained that the 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. 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 (2018d), 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.

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 re-arranged to be represented as individual-level data. Consider Exhibit 1, a two by two table with information on eight people, broken down by race and sex. It is simple to convert the table to individual-level format: we just make one record for each person. As shown in Exhibit 2, we then have exactly the same information but now expressed as microdata.

Exhibit 1. Tabular Data

	White	Black
Male	2	1
Female	3	2

Exhibit 2. Microdata

Case number	Race	Sex
1	White	Male
2	White	Male
3	White	Female
4	White	Female
5	White	Female
6	Black	Male
7	Black	Female
8	Black	Female

The first step of the Census Bureau database reconstruction experiment was an analogous rearrangement of tabular data into an individual-level format. The analysts started with a table of age by sex by race by Hispanic origin, and converted the table to microdata.[4] Then the Census Bureau added more detail on place of residence, age, and race by cross-referencing across multiple tables. To identify specific block of residence, exact age, and detailed race category, the Census Bureau turned to other tables that provided fewer dimensions but more detail on each dimension, such as single years of age or detailed race. 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 2018c). In other cases it is easy to infer individual-level variables. For example, 47% of blocks contain a single race and 60% 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% of the reconstructed cases accurately matched a case from either the swapped or the unswapped source data (Abowd 2018e; Hansen 2018). In the great majority of the mismatched cases, the errors resulted from a discrepancy in age. Given

[4] For example, if a particular census tract had three black non-Hispanic women aged 25 to 29, they could create three microdata records with these individual-level characteristics. If they repeat this process for every cell in the table, we have reconstructed the full set of microdata. The rearranged dataset contains exactly the same information as the original table, but it is represented as individual-level data.

the 50% error rate, it is not justifiable to describe the microdata as "accurately reconstructed" (Abowd 2019d).

Reconstructing microdata from tabular data does not by itself allow identification of respondents; to determine who the individuals actually are, one would then have to match their characteristics to an external identified database (including, for example, names or Social Security numbers) in a conventional re-identification attack. The Census Bureau attempted to do this by matching private data from a credit agency to the reconstructed microdata. In this attempted "putative re-identification," Census Bureau analysts matched characteristics in the reconstructed microdata to characteristics in the credit agency data. Only a small fraction of those putative re-identifications actually turned out to be correct, and the Census Bureau analysis concluded that "the risk of re-identification is small" (Abowd 2018b). An intruder would have no means of determining if any particular putative match was correct, or even to estimate the probability that a re-identification attempt succeeded. Therefore, the system worked exactly 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.

Differential Privacy for Tabular Data

Despite the failure of the individual re-identification experiment, the 100% tabular data pose some special disclosure control challenges. Because these tables include the entire population and very fine geographic detail, there could be the potential for re-identification if no disclosure protections were applied. 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. Nevertheless, it is plausible that differential privacy measures could be judiciously applied that would preserve usability for the relatively limited applications of the block-level data. The research community can help with evaluating usability for these data. If the Census Bureau makes a differentially-private version of the 2010 tabular data available to researchers, the community can evaluate usability for the needs of planners and researchers.

Differentially-private tabular data from the ACS is considerably more challenging than the 100% 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% census data. The ACS is a sample with just 1.5% of the population, 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.

Differential Privacy for 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% of the population is released to the public. There is no geographic identification of places with fewer than 100,000 inhabitants. In addition to the disclosure

protections applied to the tabular data, there are additional measures for microdata. Small categories are combined into larger ones; 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.

Despite these strong protections against re-identification, as long as microdata provide real individual-level survey responses they pose an insurmountable challenge for differential privacy. ACS microdata samples directly provide individual-level characteristics derived from real people, and this in itself represents a violation of differential privacy. Bambauer et al. (2014: 718) point out that "as Dwork herself has noted, microdata releases cannot be prepared in a way that strictly complies with differential privacy."

Senior Census Bureau scientists acknowledge that differentially private microdata will not be appropriate for investigating many research problems. 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 et al. 2018). By "*real privacy protection*," Garfinkel et al. mean differential privacy, not privacy 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 (Reiter 2019). 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 impedes new discoveries from differentially private microdata.

Abowd and Schmutte (forthcoming: 36) concur that the creation of differentially private microdata is "a daunting challenge" and argue that the best solution may be "to develop new privacy-preserving approaches to problems that have historically been solved by PUMS." They then propose that real microdata could be shielded behind an on-line query system, where the Census Bureau could define the set of allowable queries in advance. They recognize, however, that such a system would be suitable for only relatively simple analyses and propose that "More complicated analyses can be conducted in restricted-access environments" (Abowd and Schmutte forthcoming: 37).

Implications of Differential Privacy for Research

The ACS and the decennial census are among the most widely-used scientific data sources in the world. Google Scholar lists almost 70,000 references to the ACS or the 2010 census, and on average a new paper using the data appears every 30 minutes. Common topics of analysis include poverty, inequality, immigration, internal migration, ethnicity, residential segregation, disability, 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 frequently capitalize on natural experiments, leveraging discontinuities such as policy change, weather events, or earthquakes. Policy-makers and planners use small-area data from the ACS and decennial census to focus resources where they are needed. Businesses use the data to estimate future demand and determine business locations. If public use data become unusable or inaccessible because of overzealous disclosure control, there would be far-reaching consequences. The quantity and quality of research about U.S. policies, the economy, and social structure would decline precipitously.

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. There are currently 29 branches in the FSRDC network, and they have a combined total of fewer than 300 workstations. In 2018, the network is hosting a record number of 201 projects using Census data. To put this in context, IPUMS alone disseminates over 100,000 ACS and decennial census datasets to about 60,000 investigators each year (http://www.ipums.org). Other archives, such as the Inter-university Consortium for Political and Social Research, as well as the Census Bureau itself, serve hundreds of thousands of additional users. The FSRDCs do not have the capacity to handle this volume of use.

It is not easy to use the FSRDCs. Every stage of the research process is significantly more time-consuming than using public use data, and only the most persistent researchers are successful. In addition, most of the branches charge high fees for anyone unaffiliated with an institution sponsoring an FSRDC. Projects are approved only if they benefit the Census Bureau, which by itself makes most research topics ineligible. Prospective users

must prepare detailed proposals, including the precise models they intend to run and the research outputs they hope to remove from the center, which are generally restricted to model coefficients and supporting statistics. Most descriptive statistics are prohibited. Researchers are not allowed to "browse" the data or change the outputs based on their results.

Under census law, researchers must become (unpaid) Census Bureau employees to gain access to non-public data. To meet this requirement, once a project is approved researchers must obtain Special Sworn Status, which involves a level 2 security clearance and fingerprint search. Applicants must be U.S. citizens or U.S. residents for three years, so most international scholars are excluded. Researchers then undergo data stewardship training. If researchers wish to modify their original model specifications or outputs, they must submit a written request and wait for approval. When the research is complete, the results must be cleared before publication by the Center for Disclosure Avoidance Research at the Census Bureau. Any deviations from the original proposal must be documented, justified, and approved.

The FSRDCs were never intended as a substitute for public use microdata, and they cannot fulfill that role. Even if the number of seats in the centers could be multiplied several hundred-fold to accommodate the current number of users of public use data, substantial hurdles remain. Applying for access and gaining approval to use the FSRDC takes at least six months and usually more. Eligibility for using FSRDCs is limited to investigators (a) affiliated with an FSRDC (or with significant financial resources), (b) with sufficient time to wait for review and approvals, and (c) doing work deemed valuable by the Bureau. The user registration logs for the IPUMS data extraction system suggest that a

majority of census data users are graduate students, who usually lack the time and resources needed to access an FSRDC. By constraining student access to high-quality data, the proposed changes to public census and ACS data will have profound implications for the training of the next generation of social science and policy researchers.

One possibility to expand the capacity of the FSRDC network would be to create a virtual data enclave enabling online access to Census Bureau servers (Abowd 2018d). By itself, however, that approach would not solve the problem. Under Title 13, any datasets that do not fully protect confidentiality may only be used by Census Bureau employees. Thus, users of the virtual enclave, just like the current users of the FSRDCs, would have become sworn Census Bureau contractors, and all research they conduct would have to benefit the Census Bureau. Just like for the FSRDCs, all results would have to go through full disclosure avoidance review. Accordingly, even if virtual enclaves expanded the limited physical capacity of the existing FSRDC networks, current bottlenecks would only worsen with increased volume. These bottlenecks include vetting projects for feasibility and benefit to the Census Bureau, processing special sworn status applications, and reviewing outputs to approve them for release. Without a huge expansion of funding to support this work, the system would grind to a halt if tens of thousands of new researchers were suddenly added.

Privacy and Perception

Privacy in America is indeed under assault. Private companies are amassing an unprecedented volume of data about their customers, and these companies are not always good stewards of that information. Huge data breaches exposing private information about millions of people—often including credit card information and Social Security numbers—

occur with alarming frequency. There are hundreds of web sites on the Internet promising full investigative reports on any individual—including credit ratings, property records, marital history, criminal history and other information—for a modest fee. In this environment, consumers and citizens are understandably concerned about privacy.

Given this wealth of information readily available from private sources, a prospective adversary would have little incentive to turn to statistical public use data in an attempt to uncover uncertain, imprecise, and outdated information about a particular individual. The unblemished record of privacy protection in decennial census and ACS data, however, also reflects highly-effective evidence-based disclosure control measures implemented by the Census Bureau over the past five decades.

There is cause for worry, however, if there is a *public perception* that government data are not adequately safeguarded. To the extent that the public believes their responses are not truly confidential, the cooperation of respondents will probably decline. The recent revelation that Trump administration officials suggested sharing census responses with law enforcement in direct violation of census law, together with a planned new question on citizenship, are likely to have profound consequences for public trust in the Census Bureau (Brahampour 2018). Eliminating reliable public use data, however, would have little or no impact on public confidence in Census Bureau confidentiality protections. The evidence suggests that the public is not concerned about access of the academic and policy communities to public-use scientific data, so curtailing that access would not help solve the problem.

To get a sense of the nature of concerns about privacy, we searched for items related to census privacy in six newspapers published from 1938 to 2012.[5] We identified 221 news articles, letters to the editor, and opinion pieces that address the issue. Interestingly, concerns about census-related privacy have generally been declining, not increasing; by far the largest number of items date from the period around the 1940 census, when questions on income and educational attainment were first introduced. Of the 221 items, only one article and two letters cited a risk of public disclosure of census information, and those items related to the 1940 census. The other items mainly objected to government snooping on citizens, with some expressing concern about the potential for government misuse of the data for tax collection or other purposes.

The two census periods with the fewest complaints about census privacy in newspapers were 1990 and 2010. There was a modest spike in newspaper coverage around the time of the 2000 census, when Senator Lott, Senator Hagel, and Governor and presidential candidate George Bush objected to the census long form on grounds of privacy. They did not cite a risk of public disclosure; rather, in Hagel's words, "I don't know why the government needs all that information ... It's none of their damn business" (Wegner 2000).

The deluge of new discoveries attributed to census data is widely reported in the media. If that stream of research dries up, justifying the expense and burden of collecting the data will become more difficult. Handicapping researchers and planners by withholding reliable data is unlikely to allay public distrust of government data collection efforts. Indeed, when it comes to public perception, restricting reliable data to an elite set of census-

[5] *Atlanta Daily World*, *Chicago Tribune*, *Los Angeles Times*, *Minneapolis Star Tribune*, *New York Times*, and *Washington Post*.

approved contract researchers has the potential to backfire by reinforcing resentment about government snooping.

Discussion

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 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. 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.

We have six decades of precedent, reaffirmed thousands of times by the Census Bureau Disclosure Review Board, about the meaning of census law. It is well established that as long as there is great uncertainty about the *identity* of respondents, the Census Bureau may publish data that reveals the *characteristics* of respondents. That means an outsider must be unable to tie a particular record to a particular individual and have high confidence that the link is correct.

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 (forthcoming) have observed, there is a tradeoff 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, it cannot be used to optimize the privacy/usability tradeoff.

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. We need comprehensive data on work to prepare for a transformation of the nation's occupational distribution brought about by technological change. We need data to identify causes, consequences, and solutions to the opioid epidemic and declining life expectancy. To meet these varied and urgent challenges, we must have broad access to the best possible data. 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" (U.S. 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. Reliable and freely-accessible data are the bedrock of science. 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.

Recommendations

We have three specific recommendations:

1. **Differential privacy might be feasible for some tabular data, but more testing is needed before final decisions are taken.**

   The most plausible use of the technique is for the 100% 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.

2. **Differential Privacy is not appropriate or feasible for ACS microdata.**

   Multiple Census Bureau papers have reiterated the point that differentially private microdata are not appropriate for many original research problems. There is no legal mandate for differentially-private microdata. Restricted access solutions are not practical.

3. **For all data products, 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. 2017. "Research Data Centers, Reproducible Science, and Confidentiality Protection: The role of the 21st Century Statistical Agency" https://www2.census.gov/cac/sac/meetings/2017-09/role-statistical-agency.pdf

Abowd, John. 2018a. "Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau" https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html

Abowd, John. 2018b. "The U.S. Census Bureau Adopts Differential Privacy." Presented at 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining London, United Kingdom, August 23, 2018

Abowd, John. 2018c. "Staring-Down the Database Reconstruction Theorem." Presented at the Joint Statistical Meetings, Vancouver, BC, Canada, July 30, 2018

Abowd, John. 2018d. How Modern Disclosure Avoidance Methods Could Change the Way Statistical Agencies Operate. Federal Economic Statistics Advisory Committee, December 14 2018.

Abowd, John. 2018e. Personal Communication, 12/11/18.

Abowd, John and Ian Schmutte. Forthcoming. "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices." *American Economic Review*. arXiv:1808.06303v1 [cs.CR] 20 Aug 2018

Alexander, J. Trent, Michael Davern and Betsey Stevenson. 2010. "Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications." *Public Opinion Quarterly* 74 (3): 551–569,

Anderson, Margo J. and William Seltzer. 2007, "Challenges to the Confidentiality of U.S. Federal Statistics, 1910-1965," *Journal of Official Statistics,* 23, 1, 1-34.

Anderson, Margo J. and William Seltzer. 2009. "Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues." *Journal of Privacy and Confidentiality* 1: 7-52.

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): 701-755.

Barabba, Vincent. 1975. "The Right of Privacy and the Need to Know," in U.S. Census Bureau, The Census Bureau: A Numerator and Denominator for Measuring Change, Technical Paper 37. Washington, D.C.: Government Printing Office.

Brahampour, Tara. 2018. “Trump Administration Officials Suggested Sharing Census Responses with Law Enforcement, Court Documents Show.” *Washington Post*, Nov. 19.

Chicago Tribune 1910. “Taft Announces New Census.” March 15 1910, p. 4.

Cleveland, Lara, Robert McCaa, Steven Ruggles, and Matthew Sobek. 2012. “When Excessive Perturbation Goes Wrong and Why: IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata.” In Josep Domingo-Ferrer and I. Tinnirello, eds., *Privacy in Statistical Databases*. Berlin and Heidelberg: Springer Verlag, pp. 179-187.

Courtland, Sherry. 1985. “Census Confidentiality: Then and Now.” *Government Information Quarterly* 2(4): 407-418.

Dalenius, Tore. 1977. “Towards a methodology for statistical disclosure control.” *Statistik Tidskrift* 15: 429–222.

Dwork, Cynthia 2006. “Differential Privacy.” In Michele Bugliesi, Bart Preeel, Vlarimiro Sassone, and Ingo Wegener (eds) Automata, Languages and Programming. 33rd International Colloquium, ICALP 2006. Proceedings, Part II, pp. 1-12. Heidelberg: Springer.

Dwork Cynthia, Frank McSherry, Kobby Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S., Rabin T. (eds) Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, vol 3876. Heidelberg: Springer

Federal Committee on Statistical Methodology. 2005. Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22.

Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. “Issues Encountered Deploying Differential Privacy.” WPES'18 Proceedings of the 2018 Workshop on Privacy in the Electronic Society, pp. 133-137. https://dl.acm.org/citation.cfm?id=3268949

General Accounting Office. 1998. “Decennial Census: Overview of Historical Census Issues” (Chapter Report, 05/01/98, GAO/GGD-98-103).

Hansen, Mark. 2018. To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data. New York Times, Dec. 5., p. B7.

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

McClure, David and Jerome Reiter. 2012. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." *Transactions on Data Privacy,* 5: 535-552.

Reiter, Jerome P. 2019. "Differential Privacy and Federal Data Releases." *Annual Review of Statistics and Its Application*, 6:13.1–13.17.

Robbin, Alice. 2001. "The Loss of Personal Privacy and its Consequences for Social Research." *Journal of Government Information*, 28(5), 493-527.

Ruggles, Steven. 2000. "A Data User's Perspective on Confidentiality." *Of Significance: A Topical Journal of the Association of Public Data Users* 2: 1-5.

Ruggles, Steven et al. 2000. "The Public Use Microdata Samples of the U.S. Census: Research Applications and Privacy Issues." A report of the Task Force on Census 2000, Minnesota Population Center and Inter-University Consortium for Political and Social Research Census 2000 Advisory Committee. http://users.hist.umn.edu/~ruggles/Articles/2000PUMSReport.pdf

U.S. Census Bureau. 1962. Census of population and housing, 1960 public use sample: one-in-one-thousand sample. Washington, D.C.: U.S. Government Printing Office.

U.S. Census Bureau 1964. U.S. Census of Population: 1960. Availability of Published and Unpublished Data. Washington: Bureau of the Census.

U.S. Census Bureau 1967a. Data Access Descriptions. Census Tabulations Available of Computer Tape Series, CT1. December 1967. Washington: Bureau of the Census.

U.S. Census Bureau. 1967b. Small-Area Data Activities. Vol 2, no. 1 (September 1967). Washington: Bureau of the Census.

U.S. Census Bureau. 2001. A Monograph on Confidentiality and Privacy in the U.S. Census. https://www.census.gov/history/pdf/ConfidentialityMonograph.pdf

U.S. Census Bureau. 2003. 2000 Census of Population and Housing. Census 2000 Summary File 1. Washington: Bureau of the Census.

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

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

U.S. Census Bureau. 2018c. "Strategic Plan—Fiscal Year 2018 Through Fiscal Year 2022." https://www.census.gov/content/dam/Census/about/about-the-bureau/PlansAndBudget/strategicplan18-22.pdf

Wegner, Mark. 2000. "Privacy Concerns Embroil 2000 Census." Congress Daily, March 29, 2000. http://www.govexec.com/dailyfed/0300/032900b2.htm.

Wright, Caroll D. and William C. Hunt. 1900. *History and Growth of the United States Census*. Washington, D.C.: U.S. Government Printing Office.

Zayatz, Laura, et al. 1999. Disclosure Limitation Practices and Research at the U.S. Census Bureau, U.S. Bureau of the Census.