UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Risk assessment procedures for the 2020 U.S. census.

Steven Ruggles, Lara Cleveland and David Van Riper (University of Minnesota)

ruggles@umn.edu; clevelan@umn.edu; vanriper@umn.edu

Abstract

The U.S. Census Bureau plans a new approach to disclosure control for the 2020 census that will add noise to every statistic the agency produces for geographic units below the state level. The Bureau argues the new approach is needed because the confidentiality of census responses is threatened by "database reconstruction," a technique for inferring individual-level responses from tabular data. The Census Bureau constructed hypothetical individual-level census responses from public 2010 tabular data. In particular, the Bureau attempted to infer the age, sex, race, and Hispanic or Non-Hispanic ethnicity for every individual in each of the 6.3 million inhabited census blocks in the 2010 census. Using 6.2 billion statistics from nine tables published as part of the 2010 census, the Census Bureau constructed a system of simultaneous equations consistent with the published tables, and solved the system using the Gurobi linear programming software. The Census Bureau assessed the quality of the reconstruction by calculating the percentage of matches between the reconstructed data and the actual individual-level responses for each block as enumerated in the Census. The reconstructed cases matched on the four characteristics for at least one person enumerated on the same block in 46% of cases.

A major limitation of this approach is that one would expect many matches to occur purely by chance. We use a simple Monte Carlo simulation to assess how many matches would be expected randomly. Because of high residential segregation, most blocks were highly homogeneous with respect to race and ethnicity, so those characteristics are usually easy to guess. We found that if we assign age and sex randomly and then assign everyone on each block the modal race and ethnicity of each block, we obtain a match rate very close to the Census Bureau's database reconstruction. Thus, the great majority of matches reported by the Census Bureau would be expected to occur purely by chance. The match rate is therefore a poor indicator of the effectiveness of the database reconstruction. Because the database reconstruction was not substantially much more accurate than random assignment, we conclude that it does not represent a credible threat to confidentiality.

Risk Assessment Procedures for the 2020 U.S. Census

Steven Ruggles, Lara Cleveland, and David Van Riper
Institute for Social Research and Data Innovation
University of Minnesota
Email: ruggles@umn,edu

September 28, 2021

Abstract

The U.S. Census Bureau plans a new approach to disclosure control for the 2020 census that will add noise to every statistic the agency produces for geographic units below the state level. The Bureau argues the new approach is needed because the confidentiality of census responses is threatened by "database reconstruction," a technique for inferring individual-level responses from tabular data. The Census Bureau constructed hypothetical individual-level census responses from public 2010 tabular data. In particular, the Bureau attempted to infer the age, sex, race, and Hispanic or Non-Hispanic ethnicity for every individual in each of the 6.3 million inhabited census blocks in the 2010 census. Using 6.2 billion statistics from nine tables published as part of the 2010 census, the Census Bureau constructed a system of simultaneous equations consistent with the published tables, and solved the system using the Gurobi linear programming software. The Census Bureau assessed the quality of the reconstruction by calculating the percentage of matches between the reconstructed data and the actual individual-level responses for each block as enumerated in the Census. The reconstructed cases matched on the four characteristics for at least one person enumerated on the same block in 46% of cases.

A major limitation of this approach is that one would expect many matches to occur purely by chance. We use a simple Monte Carlo simulation to assess how many matches would be expected randomly. Because of high residential segregation, most blocks were highly homogeneous with respect to race and ethnicity, so those characteristics are usually easy to guess. We found that if we assign age and sex randomly and then assign everyone on each block the modal race and ethnicity of each block, we obtain a match rate very close to the Census Bureau's database reconstruction. Thus, the great majority of matches reported by the Census Bureau would be expected to occur purely by chance. The match rate is therefore a poor indicator of the effectiveness of the database reconstruction. Because the database reconstruction was not substantially much more accurate than random assignment, we conclude that it does not represent a credible threat to confidentiality.

1. Overview of Census Bureau disclosure control

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

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

The traditional Census Bureau disclosure control strategy has focused on ensuring that the identity of respondents—such as their name, address, or Social Security number—cannot be inferred from census publications. The Census Bureau implemented targeted strategies to prevent re-identification attacks so that an outside adversary cannot positively identify which person provided a particular response. The protections in place from 1970 through 2010—sampling, swapping, suppression of geographic information and extreme values, imputation, and perturbation—have worked extremely well to meet this standard (Lauger, Wisniewski, and McKenna 2014). Indeed, *there is not a single documented case of anyone outside the Census*

Bureau revealing the responses of a particular identified person using data from the decennial census.[1]

Despite the proven effectiveness of traditional statistical disclosure control, the Census Bureau adopted an entirely new methodology for disclosure control for the 2020 census based on differential privacy. Implementations of differential privacy generally involve calculating cross-tabulations from "true" data and injecting noise drawn from a statistical distribution into the cells of the cross-tabulation. There are two significant consequences of this approach:

- The noise introduced into each cell is independent of the original value of the cell. Therefore, even if the noise is small relative to the average cell value, distortions in small cell values are often proportionally large. For example, the error introduced in the population of small towns can be proportionally large, sometimes exceeding 100% of the town's true population.
- Simple random noise can produce logical inconsistencies, such as negative population counts or household counts that exceed population counts. If the data producer wishes to maintain logical consistency or preserve some noise-free counts, they must use a post-processing algorithm to adjust totals after noise injection, and this post-processing introduces additional types of error and systematic biases. In the preliminary Census Bureau demonstration datasets using differential privacy, such systematic biases are ubiquitous.[2]

[1] In recent court testimony, the Census Bureau denied our assertion that there is no documented case of outsiders identifying the responses of a particular identified person (see Defendants' Responses to Plaintiffs' First Request for Admissions, no. 6). They do not, however, document any such case of disclosure. In their sole justification for the denial, the defendants cite McKenna (2019), "U.S. Census Bureau Reidentification Studies." That citation is puzzling, since McKenna does not describe any reidentification attempts conducted outside the Census Bureau. Moreover, McKenna does not discuss any attempted reidentification of decennial census data. McKenna does describe an attempted attack on the American Community Survey, which concluded that just 0.005% of the population was vulnerable to identification. The great majority—78%—of the attempted identifications, however, were incorrect, and no identifications could be confirmed without access to the internal confidential data. McKenna's discussion therefore supports the statement that there is not a single documented case of anyone outside the Census Bureau uncovering the responses of a particular identified person using either the Decennial Census or the American Community Survey.

[2] To enable the research community to assess the consequences of differential privacy for the research and policy communities, the Census Bureau has released several demonstration datasets based on the application of differentially private algorithms to the 2010 Decennial census data. These datasets facilitate the direct comparison of published

2. The Census Bureau's database reconstruction experiment

The Census Bureau argues that differential privacy is needed because of the threat posed by database reconstruction. Database reconstruction is a process for inferring individual-level responses from tabular data (Dinur and Nissim 2003). John Abowd, the primary architect of the Census Bureau's new approach to disclosure control, argues that database reconstruction "is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared" (Abowd 2017).

Rigorous evaluation of the Census Bureau's database reconstruction experiment is important because the new approach will add error to every population count the agency produces for geographic units below the state level. Post-processing for differential privacy also introduces systematic biases in respondent characteristics that can distort the relationships among variables. Such errors and biases have the potential to significantly reduce the usability of census data for social, economic, and health research, and will compromise the integrity of basic demographic measures (Ruggles et al. 2018; Santos-Lozada et al. 2020; Hauer and Santos-Lozada 2021; Winkler et al. 2021).

Although Census Bureau staff members have repeatedly invoked database reconstruction to justify the use of differential privacy in public presentations, they have never, to our knowledge, produced a full description of their experiment, and some details remain obscure. There are no

2010 census statistics with statistics produced via the Bureau's differential private disclosure avoidance system (Van Riper et al. 2020). The Census Bureau also released the source code that had been used to implement differential privacy, enabling investigators to experiment on their own (2020 Census DAS Development Team, 2019). Over the past three years, multiple investigators seized these opportunities to understand the impact of differential privacy on census accuracy and usability. There have been several workshops and meetings devoted to the topic, including IPUMS Differential Privacy Workshop (August 15-16, 2019), the Harvard Data Science Review Symposium (October 25, 2019), the Committee on National Statistics Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations (December 11-12, 2019), and the 2020 Privacy in Statistical Databases conference (September 23-25, 2020). Additional work has appeared as working papers, as well as a few early publications (e.g., Santos-Lozada et al. 2020; Hauer and Santos-Lozada 2021; Winkler et al. 2021). The following discussion draws on insights of this research.

peer-reviewed publications explaining their methodology, and the experiment has not been replicated by outside experts. Prior to April 2021, the Census Bureau's database reconstruction experiment was documented solely in tweets and PowerPoint slides that provided few details, so it was difficult for outsiders to evaluate. In conjunction with recent legal proceedings, the Census Bureau's chief scientist has now released a more detailed description of the experiment (Abowd 2021a), and this opens new opportunities to appraise the results.

The Census Bureau's detailed description of database reconstruction provides overwhelming evidence that the database reconstruction experiment failed to demonstrate a realistic disclosure risk. On the contrary, the database reconstruction exercise provides compelling evidence that even with a massive investment of time, resources, and computing power, it would be impossible for an outside attacker to infer the characteristics of a particular individual respondent from the published tabulations used for the 2010 census.

The Census Bureau database reconstruction experiment attempted to infer the age, sex, race, and Hispanic or non-Hispanic ethnicity for every individual in each of the 6.3 million inhabited census blocks in the 2010 census. Using 6.2 billion statistics from nine tables published as part of the 2010 census, the Census Bureau constructed a system of simultaneous equations consistent with the published tables, and solved the system using Gurobi linear programming software (Abowd 2021a). This experiment provides the primary justification for the Census Bureau's adoption of differential privacy.

According to Abowd (2018a), the experiment confirmed that the individual-level census data "can be accurately reconstructed" using the published tabular census data. We contest that

assertion because the Census Bureau found that *for most of their hypothetical population,*[3] *there was not a single case in the real population that matched on block, age, sex, and race/ethnicity* (Abowd 2018b). The Census Bureau found that for 46.48% of their hypothetical population, there was at least one case in the real population that matched on block, age, sex, and race/ethnicity. Thus, there was no correct match available for 53.53% of the population.

3. The role of chance in the database reconstruction experiment

The Census Bureau's database reconstruction experiment is flawed because the Census Bureau never compared their results with a null model to evaluate how effectively it worked. As it stands, the Census Bureau experiment is like a clinical trial with no control group; just because some patients recover, that does not provide evidence that the treatment was effective. To evaluate the database reconstruction experiment, it is not sufficient to count the matches between the reconstructed population and the real population. Rather, we must assess how much the reconstruction experiment outperforms a null model of random guessing.

It is reasonable to expect one would get a lot of matches between the reconstructed data and the real data purely by chance. The Census Bureau's new documentation of the experiment shows that the "exact match rate" was positively associated with the number of people on the block (Abowd 2021a: 4): The larger the block, the more exact matches; in fact, large blocks had three times the match rate of small blocks. Database reconstruction ought to work best with small blocks where the published tables directly reveal unique combinations of respondent characteristics. The

[3] The "reconstructed" data produced by the experiment consists of rows of data identifying the age, sex, and race/ethnicity for each person in a hypothetical population of each census block; it does not include identifying information such as name, address, or Social Security number. Thus, for example, the hypothetical population of a given block could include a 26-year-old non-Hispanic white female.

obvious explanation is that larger blocks have higher odds of including by chance any specific combination of age, sex, race, and ethnicity.

The Census Bureau did not calculate the odds that they could get matches between their hypothetical reconstructed population and the actual population purely by chance. Our analysis suggests, however, that among the minority of cases where the Census Bureau did find a match between their hypothetical population and a real person, most of the matches would be expected to occur by chance.

To investigate the issue, we conducted a simple Monte Carlo simulation. We concluded that randomly chosen age-sex combinations would match someone on any given block 52.6% of the time, assuming the age, sex, and block size distributions from the 2010 census.[4] We would therefore expect the Census Bureau to be "correct" on age and sex most of the time even if they had never looked at the tabular data from 2010 and had instead just assigned ages and sexes to their hypothetical population at random.

This calculation does not factor in race or ethnicity, but because of high residential segregation, most blocks are highly homogenous with respect to race and ethnicity. If we assign everyone on each block the most frequent race and ethnicity of the block using data from the census (U.S. Census Bureau 2012), then race and ethnicity assignment will be correct in 77.8% of cases. Using that method to adjust the random age-sex combinations described above, 40.9% percent of cases would be expected to match on all four characteristics to a respondent on the same

[4] To estimate the percentage of random age-sex combinations that would match someone on a block by chance, the model generated 10,000 simulated blocks and populated them with random draws from the 2010 single-year-of-age and sex distribution. The simulated blocks conformed to the population-weighted size distribution of blocks observed in the 2010 census. The analysis then randomly drew 10,000 new age-sex combinations and searched for them in each of the 10,000 simulated blocks. In 52.6% of cases the analysis found someone in the simulated block who exactly matched the random age-sex combination.

block. That does not differ greatly from the Census Bureau's reported 46.48% match rate for their reconstructed data (Abowd 2021a: 3).

Despite the Census Bureau's massive investment of resources and computing power, the database reconstruction technique does not perform much better than a random number generator combined with a simple assignment rule for race and ethnicity. This is analogous to a clinical trial in which the treatment and the placebo produce virtually the same outcome.

4. The reidentification experiment

The Census Bureau took their experiment one step further by assessing whether their hypothetical population shared characteristics with people who appeared in non-census sources. Within each block they matched the age and sex of persons in the hypothetical population to the age and sex of persons in financial and marketing data purchased from commercial vendors after the 2010 census (Rastogi and O'Hara 2012). A match on race or ethnicity was not required for this experiment. In most cases, the hypothetical individuals constructed by the Census Bureau did not share the same age, sex, and block as anyone in the commercial data; in just 45% of cases was there at least one person in the commercial data who matched the age, sex and block number of at least one row of the hypothetical database (Abowd 2021a). This 45% match rate between the reconstructed data and the commercial data is substantially lower than one would expect by chance. Our simulation exercise—also based only on age and sex—suggests that one would expect a 52.6% match rate for a random population.

Among the cases where there was at least one person in the commercial database who matched the age, sex, and block of a row in the hypothetical population, the Census Bureau then harvested the names from the commercial database and attempted to match them with names on the same block as enumerated in the 2010 census. They found that 38% of the names from the

commercial database were actually present on the block. Based on this exercise, the Census Bureau claimed to have successfully "re-identified" 16.85% (38% of 45%) of the population (Abowd 2021a).

Once again, there is no null model for comparison purposes. One would expect that people recorded as residing on any given block in a 2010 commercial database would have a high chance of also appearing on the same block in the 2010 Census. Is the 38% match rate on names between the commercial database high or low? Without access to internal Census data, it is impossible for us to construct a usable control group, but it would have been simple for the Census Bureau to do so. In particular, the Census Bureau could have attempted to match the names of people randomly selected from the commercial database to persons in the 2010 census living on the same census block, without any reference to the Census Bureau's database reconstruction. If the 38% match rate on names for the reconstructed population is no higher than the match rate for a randomly selected subset of the commercial data, it would mean the database reconstruction has no effect on reidentification risk. Without any comparison to a null model, the match rates quoted by the Census Bureau between the commercial database and the census enumeration are not meaningful.

Reidentification means confirming the identity of a particular individual and revealing their characteristics without reference to non-public internal census files. It would be impossible to positively identify the characteristics of any particular individual using the database reconstruction without access to non-public internal census information. Abowd (2018b) acknowledged that the database reconstruction experiment demonstrates that "the risk of re-identification is small." Abowd has now retracted that statement (Abowd 2021a ¶ 83), but his supervisor has not. Acting Director of the Census Bureau Jarmin actually went farther than Abowd, writing "The accuracy of the data our researchers obtained from this study is limited, and confirmation of re-identified

responses requires access to confidential internal Census Bureau information … an external attacker has no means of confirming them" (Jarmin 2019).

5. Small blocks and swapping

In a recent supplemental court filing, the Census Bureau argues that even if most of the matches would be expected by chance, people in very small blocks are at high risk of database reconstruction (Abowd 2021b). On blocks with fewer than ten people, the Census Bureau's database reconstruction match rate for age, sex, race, and ethnicity was just over 20%, meaning that the error rate was just under 80%. Although this success rate seems low, random assignment is even worse for very small blocks; our random simulation guessed age and sex correctly in just 2.6% of cases for blocks with fewer than ten people.

The key table powering the database reconstruction experiment—Summary File 1 P012A-I—provides information on age by sex by race by ethnicity. This table can easily be rearranged into individual-level format, providing the age, sex, and race/ethnicity of the population of each block with near-perfect accuracy (Ruggles et al. 2018). How is it possible, then, that the Census Bureau's database reconstruction incorrect in almost 80% of cases? The main challenge is that that the ages in Table P012A-I are given in five-year groups instead of exact years. A random number generator would guess the correct exact age within the five-year age group approximately 20% of the time, which is very close to the accuracy level achieved by the database reconstruction experiment.

Another possible explanation for the nearly 80% error rate in the reconstruction of small blocks, as suggested in Census Bureau testimony (Abowd 2021a), is that traditional methods of disclosure control may actually be effective at protecting persons in the smallest blocks. The most

important of these methods is swapping, in which a small fraction of households are exchanged with nearby paired households that share key characteristics (McKenna 2018).

The Census Bureau recently reported on a new experiment to assess the impact of swapping on their database reconstruction experiment (Hawes and Rodriguez 2021). To simulate an extreme level of swapping, the Bureau designed an algorithm with far higher high levels of swapping and perturbation than are ever used for disclosure control. In particular, the experiment "perturbed" household size for 50% of cases and tract location in 70% of cases, and then swapped 50% of the households with someone in a different census block. In other words, they eliminated the real characteristics of the population for half the cases on each block. Then they ran the database reconstruction attack on the altered data and found that eliminating half the real population has little impact of the rate of reidentification. In this experiment, they found a match rate of age, sex, race, and ethnicity of 44.6% using unswapped data, and 42.7% on the extremely swapped data.

The Census Bureau interpreted these results to mean that even extreme swapping does not protect from database reconstruction, so differential privacy is essential. A much more plausible explanation is that the great majority of matches occurred entirely by chance, so the match rate is unaffected by substituting the data. It is likely they would get virtually the same result if instead of 50% they used a 100% swapping rate, which would mean that zero of the reidentifications would be true. Without a null model for comparison, this kind of experiment cannot be interpreted.

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

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

> No publication shall be made by the Census Office whereby the data furnished by any particular establishment or individual can be identified, nor shall the Director

> of the Census permit anyone other than the sworn employees to examine the individual reports (Reapportionment Act of 1929, CR 28 § 11).

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

For the past nine decades, the Census Bureau has interpreted the law to mean that Census Bureau publications must protect the identity of respondents. In 2002, this interpretation was codified in the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), which explicitly defined the concept of identifiable data: it is prohibited to publish "any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means" (Title 5 U.S.C. §502 (4), Public Law 107–347).

We have nine decades of precedent, reaffirmed thousands of times by the Census Bureau Disclosure Review Board, reinforcing the interpretation that the Census Bureau is prohibited from publication of statistics that disclose respondent identities. This means that an outsider cannot infer the response of a particular individual, match that response to another database, and have high confidence that the link is correct.

The disclosure controls that have been introduced over the past half-century are limited to attributes and circumstances that pose a disclosure risk through reidentification. Unlike traditional statistical disclosure control, differential privacy attempts to mask all characteristics, not just individual identities.

The Census Bureau justifies differential privacy through a novel interpretation of census law. According to Abowd (2019), "Re-identification risk is only one part of the Census Bureau's statutory obligation to protect confidentiality. The statute also requires protection against exact

attribute disclosure." Under this interpretation, the Census Bureau must not only mask the *identities* (*e.g.*, names, addresses) of respondents, but also their *characteristics* (*e.g.*, age, sex, race, or ethnicity). Abowd (2019: 16-18), argues in particular that because the 2010 census published the exact number of people of voting age in each census block, that was an exact attribute disclosure and therefore prohibited.[5]

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

Differential privacy is oriented to the protection of attributes, not the protection of identities. Accordingly, differential privacy perturbs every attribute tabulated by the census, not just the attributes that pose a risk of enabling re-identification. Because differential privacy focuses on concealing individual characteristics instead of protecting respondent identities, it is a blunt and inefficient instrument controlling disclosure of identities.

7. Conclusion: Differential privacy is a poor fit for the protection of census data

According to the Census Bureau's chief scientist Abowd (2021a: 18) "the results from the Census Bureau's 2016-2019 research program on simulated reconstruction-abetted re-

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

identification attack were conclusive, indisputable, and alarming." Abowd contends that the published tabulations of the 2010 Census "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," and with access to better commercial data an attacker "could accurately re-identify around 179 million Americans or around 58% of the population." (Abowd 2021a: 18).

Without a control group for comparison, the results reported by the Census Bureau from the database reconstruction experiment are not meaningful. As our analysis demonstrates, the threat posed by the reconstruction to respondents' confidentiality is similar to the threat posed by randomly guessing their characteristics. If a clinical trial showed that 17% of a treated population recovers, that would not prove the treatment is effective; we would also need to compare the recovery rate of a control group. Without a null model, the Census Bureau experiment fails to demonstrate that reconstruction of the tabular data poses a significant disclosure risk.

The Census Bureau's database reconstruction and reidentification exercises demonstrates that it is not plausible that an external attacker could use census tabulations to uncover the characteristics of a particular individual, for three reasons:

- The reconstructed data are usually incorrect.
- The reconstructed data usually do not match even the block, age and sex of anyone identified in outside commercial sources.
- In the minority of cases where a hypothetical reconstructed individual does match the block, age, and sex of someone in the commercial data, it usually turns out that the person identified in the commercial data was not actually enumerated on that block in the census.

Thus, the legacy disclosure control system worked exactly as intended. An outside attacker could not use database reconstruction to uncover the characteristics of a particular individual.[6]

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

In addition, there is no evidence that differential privacy reduces disclosure risk compared with traditional methods of statistical disclosure control, and it may well increase the risk. The core metric of privacy loss used in differential privacy is epsilon (ε), which is often referred to as the privacy budget. When ε is large, noise infusion is limited and privacy is low, and when ε is small, noise infusion is large, and privacy is high. It has long been recognized, however, that there is no direct relationship between the level of ε and the risk of disclosing identities. Indeed, McClure and Reiter (2012) demonstrated that the level of ε does not determine the level of disclosure risk.

[6] Abowd (2021a) App'x B ¶ 24 maintains that in a worst-case scenario (where an external attacker had data that was exactly as accurate and complete as the Census Bureau's internal data) an attacker might be able to guess a respondent's race and ethnicity and be correct in 58% of cases. This statement is inaccurate for the reasons we have detailed. It is worth noting, however, that such an exercise would be pointless even if database reconstruction did work as advertised. One could more accurately guess anyone's race and ethnicity just by assigning the most frequent race and ethnic group on the block; that guess would be correct 77.8% of the time. Calculated from U.S. Census Bureau (2011).

Because differential privacy does not target variables and circumstances that are vulnerable to attack, in some datasets with strong differential privacy (low ε), disclosure control can be weak.

The Census Bureau used ε = 19.61 for the redistricting data file. This level is many times higher than is ordinarily contemplated by privacy researchers. The range of ε in the differential privacy literature generally runs from 0.01 to 5.0, but many analysts argue that to guarantee privacy, ε should not greatly exceed 1.0 (Lee and Clifton 2011; Dwork 2011). Frank McSherry, one of the co-inventors of differential privacy, remarked that "anything much bigger than one is not a very reassuring guarantee." Criticizing Apple's use of ε = 14 to protect privacy, McSherry argued "Apple has put some kind of handcuffs on in how they interact with your data. It just turns out those handcuffs are made out of tissue paper." McSherry went on to say that "using an epsilon value of 14 per day strikes me as relatively pointless" (Greenberg 2017).

The scale of epsilon is exponential, so ε = 19.61 provides far less privacy protection than ε = 14. Accordingly, the Census Bureau's implementation of differential privacy provides an exceptionally low level of data security, and may pose greater risk than the traditional disclosure controls used by the Census Bureau.

The New York Times described a case that effectively illustrates the efficiency of traditional statistical disclosure control methods:

> The bureau has long had procedures to protect respondents' confidentiality. For example, census data from 2010 showed that a single Asian couple — a 63-year-old man and a 58-year-old woman — lived on Liberty Island, at the base of the Statue of Liberty.
>
> That was news to David Luchsinger, who had taken the job as the superintendent for the national monument the year before. On Census Day in 2010, Mr. Luchsinger was 59, and his wife, Debra, was 49. In an interview, they said they had identified as white on the questionnaire, and they were the island's real occupants.
>
> Before releasing its data, the Census Bureau had "swapped" the Luchsingers with another household living in another part of the state, who matched them on some key questions. This mechanism preserved their privacy, and kept summaries like

> the voting age population of the island correct, but also introduced some uncertainty into the data. (Hanson 2018).

Because the couple lived on a census block with only two residents, the Census Bureau recognized that they were at high risk of reidentification and thus targeted them for disclosure protection. By contrast, differential privacy makes no distinctions between high-risk and low-risk cases, so it infuses noise equally across characteristics and populations. This means that to achieve a given level of disclosure control, differential privacy must introduce far more error than would be needed using traditional statistical disclosure control.

The Census Bureau's database reconstruction exercise does not simulate a realistic attack. We do not know whether realistic attacks, such as the identification of the couple on Liberty Island, would be prevented by differential privacy. Accordingly, based on the information released to date, there is no way to be sure that a differentially private census with ε=19.61 will be as secure as a census protected by traditional disclosure controls.

The evidence supports several broad conclusions:

- The statistical disclosure controls employed by the Census Bureau over the past five censuses have proven extraordinarily effective. There is not a single documented case of anyone outside the Census Bureau uncovering the responses of a particular identified person using data from the decennial census.
- The Census Bureau's database reconstruction experiment—the chief rationale for adopting differential privacy—failed to demonstrate a credible threat to the exposure of individual identities to anyone outside the Census Bureau. The Acting Director of the Census Bureau confirmed this interpretation when he wrote "The accuracy of the data our researchers obtained from this study is limited, and confirmation of re-identified responses requires access to

confidential internal Census Bureau information … an external attacker has no means of confirming them" (Jarmin 2019).

- The Census Bureau's novel contention that census law prohibits "exact disclosure of attributes" even if identities are fully masked is an obvious misinterpretation of the intent of Congress and contradicts centuries of precedent. Following every census since 1790, the Census Bureau has published exact attributes just as they were enumerated.
- At the proposed privacy budget level, there no guarantee that the Census Bureau's new approach increases protection of identities compared with traditional statistical disclosure controls; in fact, it may provide less protection.

Differential privacy is an inappropriate technique for disclosure control in the census, since it is a blunt and inefficient instrument that adds unnecessary error to every statistic, even though most statistics pose no risk of a breach of confidentiality (Domingo-Ferrer, Sánchez, and Blanco-Justicia 2020). The adoption of a new regime of disclosure protection is justified only if the benefit of increased protection of respondent identities outweighs the cost inflicted by damage to the utility of the data. The census includes just a few basic population characteristics: age, sex, race, Hispanic origin, family relationship, home ownership, and home occupancy. This information is not highly sensitive and can often be readily obtained from public sources such as voter-registration or property records. Even if database reconstruction worked as described, it is implausible that an outside attacker would invest the enormous time and resources needed to develop reconstructed individual-level census data from published tabulations. Given that the database reconstruction method developed by the Census Bureau performs little better than a roll of the dice, we can be confident that malicious intruders pose no realistic threat of harm.

References

2020 Census DAS Development Team. 2019. Disclosure Avoidance System for the 2020 Census, End-to-End Release: Uscensusbureau/Census2020-Das-E2e. U.S. Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e.

Abowd, John. 2017. "Research Data Centers, Reproducible Science, and Confidentiality Protection: The role of the 21st Century Statistical Agency." U.S. Census Bureau. Presentation to the Summer DemSem, June 5, 2017. https://www2.census.gov/cac/sac/meetings/2017-09/role-statistical-agency.pdf

Abowd, John. 2018a. "How Modern Disclosure Avoidance Methods Could Change the Way Statistical Agencies Operate." Federal Economic Statistics Advisory Committee, December 14 2018. https://www.census.gov/content/dam/Census/about/about-the-bureau/adrm/FESAC/meetings/Abowd%20Presentation.pdf

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

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

boyd, dana. 2020. "Balancing Data Utility and Confidentiality in the 2020 US Census." Version of April 27, 2020. https://datasociety.net/wp-content/uploads/2019/12/Differential-Privacy-04_27_20.pdf

Abowd, J. 2021a. 2010 Declaration of John Abowd, State of Alabama v. United States Department of Commerce. Case No. 3:21-CV-211-RAH-ECM-KCN. (2021) Appendix B: 2010 reconstruction-abetted re-identification simulated attack.

Abowd, J. 2021b. 2010 Supplemental Declaration of John M. Abowd, State of Alabama v. United States Department of Commerce. Case No. 3:21-CV-211-RAH-ECM-KCN. (2021)

Dajani, Aref N. et al. 2017. The modernization of statistical disclosure limitation at the U.S. Census Bureau. https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf

Dinur, I., & Nissim, K. 2003. Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 202-210.

Domingo-Ferrer, J., Sánchez, D. and Blanco-Justicia, A., 2020. "The limits of differential privacy (and its misuse in data release and machine learning)." arXiv preprint https://arxiv.org/pdf/2011.02352.pdf

Dwork, C., 2011. "A firm foundation for private data analysis." Communications of the ACM, 54(1), pp.86-95.

Garfinkel, Simson. 2017. “Modernizing Disclosure Avoidance: Report on the 2020 Disclosure Avoidance Subsystem as Implemented for the 2018 End-to-End Test.” Presentation at the Census Scientific Advisory Committee, September 17, 2017. https://perma.cc/4J8B-ZEXM

Garfinkel, Simson, John M Abowd, and Christian Martindale. 2018. “Understanding Database Reconstruction Attacks on Public Data.” ACM Queue 16 (5).

Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. “Issues Encountered Deploying Differential Privacy.” In Proceedings of the 2018 Workshop on Privacy in the Electronic Society, 133–137. WPES’18. New York, NY, USA: ACM. https://doi.org/10.1145/3267323.3268949.

Greenberg, A. 2017. How One of Apple's Key Privacy Safeguards Falls Short. Wired Magazine, 9/15/2017. https://www.wired.com/story/apple-differential-privacy-shortcomings/

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

Hauer ME, Santos-Lozada AR 2021. Differential Privacy in the 2020 Census Will Distort COVID-19 Rates. Socius. doi:10.1177/2378023121994014

Hawes, M., Rodriguez, R.A. 2021. Determining the Privacy-loss Budget Research into Alternatives to Differential Privacy. Census Bureau Webinar, May 25, 2021. https://www2.census.gov/about/partners/cac/sac/meetings/2021-05/presentation-research-on-alternatives-to-differential-privacy.pdf

Hawes, Michael and Wright, Tommy. 2021. “Summary of DAS Status and P.L. 94-171 Tuning Experiments.” Committee on National Statistics, National Academy of Sciences, Expert Groups Webinar, March 30, 2021.

Jarmin, Ron. 2019. “Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census.” Director’s Blog, U.S. Census Bureau. https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census_bureau_adopts.html

Lauger, Amy, Billy Wisniewski, and Laura McKenna. 2014. “Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research.” Research Report Series (Disclosure Avoidance #2014-02). Washington: Center for Disclosure Avoidance Research, U.S. Census Bureau. https://www.census.gov/library/working-papers/2014/adrm/cdar2014-02.html

Leclerc, Philip. 2019. “The 2020 Decennial Census TopDown Disclosure Limitation Algorithm: A Report on the Current State of the Privacy Loss–Accuracy Trade-Off.” In . Washington DC. https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518.

Lee, J. and Clifton, C., 2011, October. “How much is enough? Choosing ε for differential privacy.” In International Conference on Information Security (pp. 325-340). Springer, Berlin, Heidelberg.

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

McKenna, L. 2018. “Disclosure avoidance techniques used for the 1970 through 2010 decennial censuses of population and housing.” Technical report 18–47, US Census Bureau, Washington, DC. https://www.census.gov/library/working-papers/2018/adrm/cdar2018-01.html

McKenna, L. 2019. “U.S. Census Bureau Reidentification Studies.” Research and Methodology Directorate, U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2020/adrm/8%20T13%20Reidentification%20studies(tagged)%20CED-DA%20Report%20Series.pdf

Rastogi Sonia and O’Hara, Amy. 2012. “2010 Census Match Study.” 2010 Census Planning Memoranda Series, no. 247. U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/publications/2012/dec/2010_cpex_247.pdf

Ruggles, S., Fitch, C., Magnuson, D., Schroeder, J. 2018. Differential privacy and census data: Implications for social and economic research. AEA Papers and Proceedings 109, 403-408.

Ruggles, S. et al. 2018. Implications of differential privacy for Census Bureau data and scientific research. Minneapolis, MN: Minnesota Population Center, University of Minnesota (Working Paper 2018-6). https://assets.ipums.org/_files/mpc/wp2018-06.pdf

Santos-Lozada, A.R., Howard, J.T., Verdery, A.M. 2020. How differential privacy will affect our understanding of health disparities in the United States. PNAS 117 (24) 13405-13412.

U.S. Census Bureau. 2011. “Census 2010 Summary File1 - P5. Hispanic or Latino Origin by Race.” Retrieved from https://www.nhgis.org.

U.S. Census Bureau. 2021a. “Group Quarters Imputation Methodology.” In Case 1:21-cv-01361-ABJ, Document 8-7, pp. 258-267.

U.S. Census Bureau. 2021b. “Census Bureau Administrative Data Inventory - March 2021.” Retrieved 4/9/2021 from https://www2.census.gov/about/linkage/data-file-inventory.pdf.

Van Riper, D., Kugler, T., and J. Schroeder. IPUMS NHGIS Privacy-Protected 2010 Census Demonstration Data [Database]. Minneapolis, MN: IPUMS. 2020.

Winkler, R.L., Butler, J.L., Curtis, K.J. et al. 2021. Differential Privacy and the Accuracy of County-Level Net Migration Estimates. Population Research and Policy Review.