UNIVERSITY OF MINNESOTA

The Role of Chance in the Census Bureau Database Reconstruction Experiment

Steven Ruggles†
University of Minnesota

David Van Riper
University of Minnesota

May 2021

Working Paper No. 2021-01
DOI: https://doi.org/10.18128/MPC2021-01

†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 Alfred P. Sloan Foundation G-2019-12589, "Implications of Differential Privacy on Decennial Census Data Accuracy and Utility" and the Minnesota Population Center at the University of Minnesota (P2C HD041023).

Abstract: The Census Bureau plans a new approach to disclosure control for the 2020 census that will add noise to every statistic the agency produces for places 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 and matched them to internal census records and to outside sources. We implement a simple simulation to assess how many matches would be expected by chance. We demonstrate that most matches reported by the Census Bureau experiment would be expected randomly. The database reconstruction experiment therefore fails to demonstrate a credible threat to confidentiality.

Database reconstruction is a process for inferring individual-level responses from tabular data (Dinur and Nissim 2003). 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). 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 2021), and this opens new opportunities to appraise the results.

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 2021). This experiment provides the primary justification for the Census Bureau's adoption of differential privacy.

We argue that the database reconstruction experiment is flawed because the results reported by the Census Bureau would be expected to occur mainly by chance. This finding is important because differential privacy will add error to every statistic the agency produces for geographic units below the state level, and this error will significantly reduce the usability of census data for social, economic, and health research (Ruggles et al. 2018; Santos-Lozada et al. 2020; Hauer and Santos-Lozada 2021).

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. The Census Bureau found that for 53.52% of their hypothetical population, there was not a single case in the real population that matched on block, age, sex, and race/ethnicity. There was at least one person who matched on all characteristics in 46.48% of cases (Abowd 2021).

The Census Bureau then assessed 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 2021).

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 2021).

One would expect to get many 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 2021: 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, not large ones. The obvious explanation is that larger blocks have higher odds of including by chance any specific combination. of age, sex, race, and ethnicity.

In the real 2010 population, 57% of persons are unique at the census block-level with respect to age, sex, race, and ethnicity (Abowd 2021). This means that 43% of persons reside on a block with one or more other people who share their exact characteristics. This also suggests that a person with randomly selected characteristics would have a reasonably high chance of exactly matching someone on any given block.

The Census Bureau did not, apparently, 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 estimate 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. To estimate the percentage of random age-sex combinations that would match someone on a block by chance, we 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. We then randomly drew 10,000

new age-sex combinations and searched for them in each of the 10,000 simulated blocks.[1] In 52.6% of cases we found someone in the simulated block who exactly matched the random age-sex combination. The relationship between block size and the percent of random age-sex combinations present appears in Figure 1.

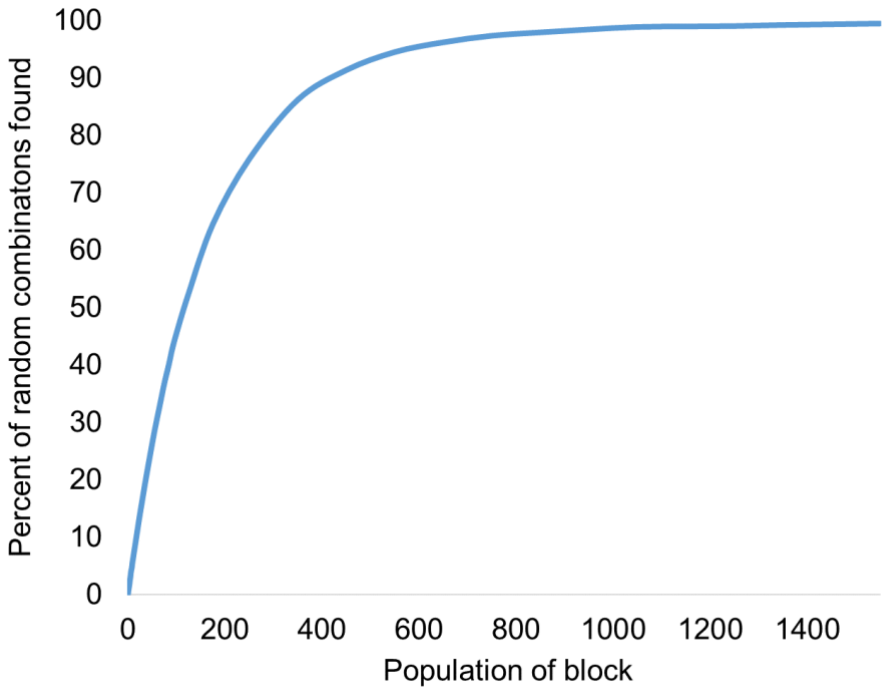

Fig. 1. Percent of randomly selected age-sex combinations present by size of block. The average person in the 2020 Census resided on a block with 249.5 people. For blocks of that size, one would expect any randomly chosen age-sex combination to be present 75.8% of the time

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. The 52.5% match rate for the random population is substantially higher than the 45% match rate that the Census Bureau found between the reconstructed data and the commercial data, which also was based just on age and sex. The

[1] Our simulation code and supporting data files are available at http://users.hist.umn.edu/~ruggles/censim.html.

randomly simulated population was similar to the real census population with respect to the frequency of unique respondents: we found that 47.7% of the simulated population was unique within the block with respect to age and sex, compared with 44% in the real population (Abowd 2021).

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

Acting Director of the Census Bureau Ron Jarmin supports the use of differential privacy, but at the same time acknowledges that the database reconstruction experiment failed to demonstrate a serious threat to the confidentiality of 2020 Census responses. He wrote that "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 "reconstructed" data is usually false, and an intruder would have no means of determining if any particular inference was true. Our simulation exercise now demonstrates that that most of the matches reported by the Census Bureau would be expected to occur purely by chance. This analysis reinforces the

conclusion that the database reconstruction experiment failed to demonstrate a credible threat to census confidentiality.

References

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

Dinur, I., & Nissim, K. (3003) Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 202-210.

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

Rastogi, S. & O'Hara, A. (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.

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. Census 2010 Summary File1 - P5. Hispanic or Latino Origin by Race (2012). Retrieved from https://www.nhgis.org.

Jarmin, R. (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