Implications of Differential Privacy for Census Bureau Data Dissemination

Steven Ruggles

Institute for Social Research and Data Innovation

University of Minnesota

December 2018

Acknowledgements

This report was prepared by Steven Ruggles (ISRDI) with the assistance of 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 (Sociology, 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 (Census Bureau), 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).

Outline

1. Brief History of Census Privacy Policies
2. Differential Privacy and Census Law
3. Challenges of Differentially-private Microdata
4. Conclusion

Outline

1. Brief History of Census Privacy Policies
2. Differential Privacy and Census Law
3. Challenges of Differentially-private Microdata
4. Conclusion

Highlights in the history of census privacy

- 1929: Census law made protection explicit

  "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."
- 1954: Title 13 retained 1929 language
- 1962: No sharing within government, immune from legal process
- 2002: Confidentiality requirements clarified by the "Confidential Information Protection and Statistical Efficiency Act" (CIPSEA) formally defined the meaning of identifiable data

1962: The first electronic data publication

- 1-in-1000 microdata sample
- Confidentiality protections: eliminating personal identifiers, low-level geography, top-coding income.
- “It has been determined that making records available in this form does not violate the provision of confidentiality under which the census was conducted”

Key developments since 1962

- 1990: Swapping and imputation
- 2000: Microdata debate and compromise
- 2018: New disclosure rules that mark a “sea change for the way that official statistics are produced and published.” (Garfinkel et al. 2018)

Outline

1. Brief History of Census Privacy Policies
2. Differential Privacy and Census Law
3. Challenges of Differentially-private Microdata
4. Conclusion

Database reconstruction

- The new disclosure rules were motivated by the threat of “database reconstruction”
- As applied by the Census Bureau this is the process of inferring individual-level data from tabular data
- According to Abowd (2017), database reconstruction “is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared.”

Database reconstruction

- Any tabular data can be expressed as microdata
- Census Bureau reconstruction experiment begins by expressing a table of age by sex by race by Hispanicity as microdata
- Using multiple tables, Census analysts inferred details on place of residence and age not available in any single table

Tabular Data

	White	Black
Male	2	1
Female	3	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

Database reconstruction experiment

- “Correctly” identifies age, sex, race, and Hispanic ethnicity for an average of 50% of persons in each block
  - Low match rate may partly reflect census confidentiality measures, especially swapping
  - Some blocks are indeterminate
- At this point, this does not rise to claim of “accurately reconstructed” or “quite accurate” microdata
- An outside attacker would have no means of determining which of the records were true

Reconstruction vs. re-identification

- Database reconstruction should not be confused with re-identification
- The reconstructed microdata have no identifying information: just block, age, sex, race, and whether Hispanic
- To identify anyone's characteristics, one would have to match the reconstructed microdata to another source that includes identifiers such as names

Census Bureau re-identification attempt was unsuccessful (which is good)

- Census Bureau analysis concluded that “the risk of re-identification is small.” (Abowd 2018)
- The disclosure control system apparently works as designed: because 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 already sufficient uncertainty to make *positive* identification by an outsider impossible

So why is database reconstruction a problem?

The concern is based on a novel reading of this clause of Title 13:

> "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))

Re-interpreting census law

- Since 1962, the Census Bureau has interpreted “any particular establishment or individual” to mean an individual whose identity can be determined
- Now some are saying the Census Bureau cannot release data about individuals, even if the identity of those individuals is unknown

Re-interpreting census law

Six decades of history and precedent, as well as the 2002 CIPSEA law, support the traditional Census Bureau interpretation of Title 13:

> The Census Bureau cannot reveal “the identity of the respondent to whom the information applies.” (Title 5 U.S.C. §502 (4))

This has been amazingly successful: There are no documented instances in which the identity of anyone in the decennial census of the ACS has been determined by anyone outside the Census Bureau.

The “death knell” for census data

- The new interpretation asserts that it is prohibited to reveal *characteristics* of an individual even if the *identity* of that individual is effectively concealed
- This is a radical departure from established census law and precedent

Special sensitivity of 100% summary files

- Even if current summary files are not in violation of census law there may be cause for concern because these are 100% data files at the block level
- DP techniques may be feasible because the use cases for the block-level short-form data are limited (mainly reapportionment, aggregation to higher levels, and residential segregation)
- Further testing is needed to evaluate whether DP block-level data will meet the needs of researchers and planners

ACS summary files are inherently less sensitive

1. It is a sample (about 1.5% of housing units annually) so it is highly unlikely any particular individual is represented in the data

   If a case is uniquely matched by characteristic to an identified dataset, there is no way to determine that the match is correct, since the true match may not have been sampled.
2. There is no block data. Smallest geography is for the block group, and those tables are very limited.
3. ACS small-area data is already very blurry; DP might not be much worse.

ACS microdata files are even more protected

- It is a sample of a sample (currently about 0.96% of the population is included annually) so it even more highly unlikely that any particular individual is represented
- Smallest geography is the PUMA, with at least 100,000 persons
- An attacker could never determine whether or not any match was actually the targeted “particular individual”
- Differential privacy is not a realistic goal for microdata; Every indication is that DP would seriously compromise usability

Outline

1. Brief History of Census Privacy Policies
2. Differences between Differential Privacy and Census Law
3. Challenges of Differentially-private Microdata
4. Conclusions

Microdata representing real individual-level responses cannot strictly comply with differential privacy

Garfinkel et al. (2018):
"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."

What this means:

- The Census Bureau can't make differentially private microdata useful for uncovering relationships that are not anticipated in advance and intentionally baked into the database
- This makes new discoveries from differentially private microdata unlikely

The proposed solution:

Garfinkel et al. (2018):
"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.

Abowd and Schmutte (forthcoming) concur:

Formally private microdata is “a daunting challenge”

Best solution may be “to develop new privacy-preserving approaches to problems that have historically been solved by PUMS.”

- Online query system, with predetermined allowable queries
- Restricted data solutions

The ACS is among the most widely-used scientific data sources in the world

Google Scholar lists 55,000 references to the ACS, and on average a new paper using the data appears every 55 minutes.

Restricted access is not a viable alternative

- The FSRDC network would have to be expanded by two orders of magnitude to accommodate demand
- Under law, only *sworn Census Bureau employees* may access data protected by Title 13
- RDC users must become Census Bureau contractors, and obtain special Sworn status
- Therefore, even if data were disseminated through a virtual data enclave, all the hoops would remain

Restricted Access is not a viable alternative

- Projects must benefit Census Bureau
- No non-U.S. residents
- Level 2 security clearance
- Existing bottlenecks would get worse
  - Obtaining Special Sworn Status
  - Project approval
  - Reviewing outputs for disclosure

Outline

1. Brief History of Census Privacy Policies
2. Differences between Differential Privacy and Census Law
3. Challenges of Differentially-private Microdata
4. Conclusions

Conclusions

1. Differential privacy may be feasible for summary files, especially the 100% census tabulations.

   To assess whether differentially private summary files will meet the needs of investigators, the research community should be given access to preliminary versions of noise injected data as soon as they are available.

Conclusions

2. Microdata is much more problematic. Multiple Census Bureau papers have reiterated the point that differentially private microdata will not be appropriate for most original research problems.

   Restricted access solutions will not work.

Conclusions

3. There is no legal mandate for differential privacy. As long as the identify of the respondents cannot be determined, public use microdata is consistent with Title 13, CIPSEA, and decades of Census Bureau precedent.

   There is compelling need for valid microdata; ACS microdata are essential for addressing critical challenges facing the United States.

Conclusions

4. The Census Bureau should continue to pursue modernization of microdata disclosure control even though differential privacy is not attainable. This may include strategic noise injection focusing on variables and subpopulations at greatest risk of re-identification.

   The research community should have an opportunity to test any new disclosure control procedures through a rigorous process, by replicating past peer-reviewed research using data with the new disclosure controls.

Census Mission

The Census Bureau's core mission is “to serve as the nation’s leading provider of quality data about its people and economy”

- The Census Bureau has extraordinary record—better than anywhere else in the world—of making powerful public use data broadly accessible
- Just as important, the Census Bureau also has a stellar record of protecting confidential information
- We must ensure that both of these powerful traditions continue

Thank You