Implications of Differential Privacy for Public Data

Steven Ruggles

Institute for Social Research and Data Innovation
University of Minnesota
September 2018: Differential Privacy Announced

• New disclosure rules mark a “sea change for the way that official statistics are produced and published.”
• Applied first to 2020, and “will then be adapted to protect publications from the American Community Survey and eventually all of our statistical releases.”
Our Concerns

• Differential privacy will degrade the quality of data available about the population, and will probably make scientifically useful public use microdata impossible

• The differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau
Differential Privacy redefines “privacy”

- DP is not a method of disclosure control
- DP is a new formal definition of privacy:

\[ \Pr [M(D_0) \in S] \leq e^{\epsilon} \cdot \Pr [M(D_1) \in S] + \delta \]

- This yields a measure called Epsilon (\(\epsilon\)) that defines the level of “privacy” in a dataset.
- A small \(\epsilon\) means high “privacy”
Census law requires protection of identities

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

• The Census Bureau’s disclosure control program has focused on protection of identities

• This has been 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
Differential privacy does not measure disclosure risk!

“disclosure risk measures can be small even when \( \epsilon \) is large.”

**Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data**

David McClure*, Jerome P. Reiter**
*Department of Statistical Science, Duke University, Box 90251, Durham, 27708, USA.
**Department of Statistical Science, Duke University, Box 90251, Durham, 27708, USA.
E-mail: david.mcclure52@gmail.com, jerry@stat.duke.edu

**Abstract.** We compare the disclosure risk criterion of \( \epsilon \)-differential privacy with a criterion based on probabilities that intruders uncover actual values given the released data. To do so, we generate fully synthetic data that satisfy \( \epsilon \)-differential privacy at different levels of \( \epsilon \), make assumptions about the information available to intruders, and compute posterior probabilities of uncovering true values. The simulation results suggest that the two paradigms are not easily reconciled, since differential privacy is agnostic to the specific values in the observed data whereas probabilistic disclosure risk measures depend greatly on them. The results also suggest, perhaps surprisingly, that probabilistic disclosure risk measures can be small even when \( \epsilon \) is large. Motivated by these findings, we present an alternative disclosure risk assessment approach that integrates some of the strong confidentiality protection features in \( \epsilon \)-differential privacy with the interpretability and data-specific nature of probabilistic disclosure risk measures.
Differential Privacy is not concerned with re-identification of respondents

- DP prohibits revealing *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
- The Census Bureau has been disseminating individual-level *characteristics* routinely since the first microdata in 1962
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.

10:33 AM - 7 Apr 2019

This is a major re-interpretation of the law. If it were correct, microdata with any real responses would be illegal, even if identities were effectively protected.
Database reconstruction experiment

- The new disclosure rules were motivated by the threat of “database reconstruction”
- According to Abowd (2017), database reconstruction “is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared.”
An attacker could use database reconstruction to guess someone’s race or Hispanic origin, but they would usually be wrong and would have no means of determining whether or not they were correct.

“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.’’

Ron Jarmin, Deputy Director
Blog Post 2/15/19
Evaluating impact of DP using the 1940 census

In June, Census provided noise-infused data from the 1940 census.

This graph shows impact of noise when $\epsilon = .25$, the same noise level as used in the 2018 dress rehearsal.
Evaluating impact of DP using the 1940 census

Each dot is a Minnesota enumeration district, about the size of block groups

Vertical axis: % adult in noise infused data
Horizontal axis: % adult in real 1940 data
Evaluating impact of DP using the 1940 census

Many of the noise-infused districts have 100% or 0% adults (100% children!).

These data would not be usable for drawing school district boundaries.
Percent aged 18+, Minnesota Enumeration Districts, comparing different levels of noise infusion (Epsilon)

- $\epsilon = 0.25$
- $\epsilon = 0.50$
- $\epsilon = 0.75$
- $\epsilon = 1.0$
- $\epsilon = 2.0$
- $\epsilon = 4.0$
- $\epsilon = 6.0$
- $\epsilon = 8.0$

Original 1940 data
Diversity Index, Minnesota Enumeration Districts, comparing different levels of noise infusion (Epsilon)

Original 1940 data

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

- To guarantee differential privacy, microdata must be simulated using statistical models rather than directly derived from the responses of real people.
- The Census Bureau can’t make differentially private microdata that useful for uncovering relationships that are not anticipated in advance and intentionally baked into the database.
Conclusions and Recommendations

1. Differential privacy may make tabular data unusable for most applications of small-area data

2. Differential privacy is not appropriate or feasible for ACS microdata

3. Alternative methods that focus on disclosure control rather than differential privacy could optimize the tradeoff between risk and usability
Thank You