Title 13, Differential Privacy, and the 2020 Decennial Census

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

Oregon State Data Center Meetings
October 1-3, 2019

Shape your future START HERE >

United States Census 2020

Presentation Overview

- **Title 13 and our commitment to data stewardship**
- **Where we came from: the Census Bureau's privacy protections over time**
- **The growing threat of re-identification**
- **Differential Privacy – what it is, and what it isn't!**
- **Implications for the 2020 Decennial Census**
- **Questions**

For more information and technical details relating to the issues discussed in these slides, please contact the author at michael.b.hawes@census.gov.

Any opinions and viewpoints expressed in this presentation are the author's own, and do not necessarily represent the opinions or viewpoints of the U.S. Census Bureau.

Shape your future START HERE >

United States Census 2020

In the News

Reconstruction and Reidentification are not just theoretical possibilities…they are happening!

- **Massachusetts Governor's Medical Records** (Sweeney, 1997)
- **AOL Search Queries** (Barbaro and Zeller, 2006)
- **Netflix Prize** (Narayanan and Shmatikov, 2008)
- **Washington State Medical Records** (Sweeney, 2015)
- and many more…

Reconstructing the 2010 Census

The 2010 Census collected information on the age, sex, race, ethnicity, and relationship (to householder) status for ~309 Million individuals. (1.9 Billion confidential data points)

The 2010 Census data products released over 7.7 Billion statistics.

Internal Census Bureau research confirms that the confidential 2010 Census microdata can be accurately reconstructed from the publicly released tabulations.

Reconstructing the 2010 Census: How did we do it?

- **Performed database reconstruction for all 308,745,538 people enumerated in the 2010 Census from public 2010 data products.**
- **Linked reconstructed records to commercially available databases.**
- **Successful record linkage to commercial data = "putative re-identification"**
- **Compared putative re-identifications to confidential data.**
- **Successful linkage to confidential data = "confirmed re-identification"**
- **Potential harm to individuals: can learn self-response race and ethnicity.**

Reconstructing the 2010 Census: What did we find?

- **Census block and voting age (18+) were correctly reconstructed in all 6,207,027 inhabited blocks.**
- **Block, sex, age (in years), race (OMB 63 categories), and ethnicity were reconstructed:**
  - Exactly for 46% of the population (142 million individuals)
  - Within +/- one year for 71% of the population (219 million individuals)
- **Block, sex, and age were then linked to commercial data, which provided putative re-identification of 45% of the population (138 million individuals).**
- **Name, block, sex, age, race, ethnicity were then compared to the confidential data, which yielded confirmed re-identifications for 38% of the putative re-identifications (52 million individuals).**
- **For the confirmed re-identifications, race and ethnicity are learned correctly, though the attacker may still have uncertainty.**

The Census Bureau's Decision

Advances in computing power and the availability of external data sources make database reconstruction and re-identification increasingly likely.

The Census Bureau recognized that its traditional disclosure avoidance methods are increasingly insufficient to counter these risks.

To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections.

Differential Privacy

aka "Formal Privacy"

-quantifies the precise amount of privacy risk…

-for all calculations/tables/data products produced…

-no matter what external data is available…

-now, or at any point in the future!

Assessing Privacy Risk

Traditional Disclosure Avoidance Considers Absolute Privacy Risk

Can an individual be re-identified in the data, and can some sensitive attribute about them be inferred?

Evaluates risk given a particular, defined mode of attack, asking: What is the likelihood, at this precise moment in time, of re-identification and inferential disclosure by a particular type of attacker with a defined set of available external information?

Formal Privacy is about Relative Privacy Risk

Does not directly measure re-identification risk (which requires specification of an attacker model).

Instead, it defines the maximum privacy "leakage" of each release of information compared to some counterfactual benchmark (e.g., compared to a world in which a respondent does not participate, or provides incorrect information).

Precise amounts of noise

Differential privacy allows us to inject a precisely calibrated amount of noise into the data to control the privacy risk of any calculation or statistic.

Privacy vs. Accuracy

Differential Privacy also allows policymakers to precisely calibrate where on the privacy/accuracy spectrum the resulting data will be.

Data Quality|Bnae Kegouqe
Dada Qualitg|Vrkk Jzcfkdy
Data Qaality|Dncb PrhvBln
Dzte Qvality|Dncb Prtnavy
Dfha Quapyti|Tgta Ppijacy
Tgta Qucjity|Dfha Pnjvico
Dncb Qhulitn|Dzhe Njivaci
Ntue Quevdto|Dzte Privecy
Vrkk Zuhnvry|Dada Privacg
Bnaq Denorbe|Data Privacy

Establishing a Privacy-loss Budget

The only way to absolutely eliminate all risk of re-identification would be to never release any usable data.

Differential privacy allows you to quantify a precise level of "acceptable risk" of re-identification.

This measure is called the "Privacy Budget" or "Epsilon."

- **ε=0 (perfect privacy) would result in completely useless data**
- **ε=∞ (perfect accuracy) would result in releasing the data in fully identifiable form**

The Formal Guarantee

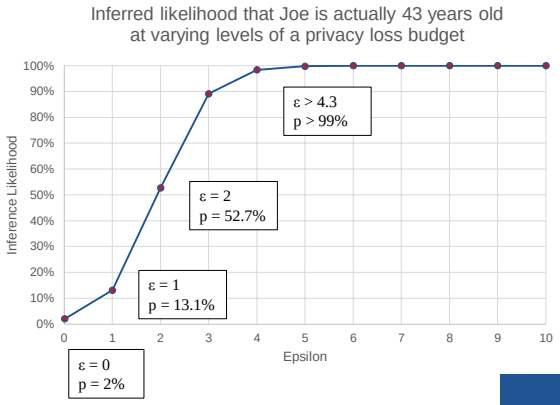
Can Sara determine Joe's exact age?

Suppose Joe submitted erroneous information for the Census, and the best Sara could otherwise do to determine Joe's exact age (from other available information) is to predict that there is a 2% chance that Joe is 43 years old.

If Joe instead provides accurate information for the Census, then a small amount of information about him will "leak" through the publication of data products. This new information can improve Sara's estimate.

The Privacy-loss Budget determines the amount of that leakage and the corresponding maximum possible improvement to Sara's prediction.

Assumes that Sara has infinite computing resources, infinitely powerful algorithms, and allows her to have arbitrary side knowledge.



Allocating the Privacy-loss Budget

Each calculation, query, or tabulation of the data consumes a fraction of the privacy-loss budget.

$(\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 \ldots + \varepsilon_n = \varepsilon_{Total})$

Calculations/tables for which high accuracy is critical can receive a larger share of the overall privacy-loss budget.

Keeping Accuracy High

When Differential Privacy is applied, the accuracy of the resulting data will be affected by:

- **The number of calculations being performed or tables being generated;**
- **The type of calculation being performed (e.g., count vs. mean);**
- **The size of the underlying populations for each calculation or table;**
- **The range of possible values;**
- **The overall privacy budget (epsilon); and**
- **The allocation of the privacy budget across calculations/tables.**

Comparing Methods

Data Accuracy

Differential Privacy is not inherently better or worse than traditional disclosure avoidance methods.

Both can have varying degrees of impact on data quality depending on the parameters selected and the methods' implementation.

Privacy

Differential Privacy is substantially better than traditional methods for protecting privacy, insofar as it actually allows for measurement of the privacy risk.

Implications for the 2020 Decennial Census

The switch to Differential Privacy will not change the constitutional mandate to reapportion the House of Representatives according to the actual enumeration.

As in 2000 and 2010, the Census Bureau will apply privacy protections to the PL94-171 redistricting data.

The switch to Differential Privacy requires us to re-evaluate the quantity of statistics and tabulations that we will release, because each additional statistic uses up a fraction of the privacy budget (epsilon).

In order to maximize the accuracy of the data, the Census Bureau is carefully evaluating what tabulations will be released at different levels of geography.

You Can Help Us to Help You!

Senior Census Bureau policymakers will be making important decisions – and they need your input!

The actual impact of Differential Privacy on the usability and accuracy of the 2020 Census data products will ultimately depend on the following factors:

- What will the overall privacy budget (epsilon) be?
- What statistics will the Census Bureau release at which levels of geography?
- How will the overall privacy budget be allocated across different geographies, tables, and products?

In order for the Census Bureau's senior leadership to make the most informed decisions on these questions, they need to know how you plan to use the 2020 Census data.

2010 Demonstration Products

- Census Bureau has released a set of data products that demonstrate the computational capabilities of the DAS. The current version of the DAS was run on the 2010 internal data to produce two products:
  - PL 94-171
  - Demographic and Housing Characteristics File (selected tables)
- Allows data users to assess the impacts of the DAS implementation.

Available at: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html

Questions?

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

301-763-1960 (Office)
michael.b.hawes@census.gov

Shape your future START HERE >
United States Census 2020