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

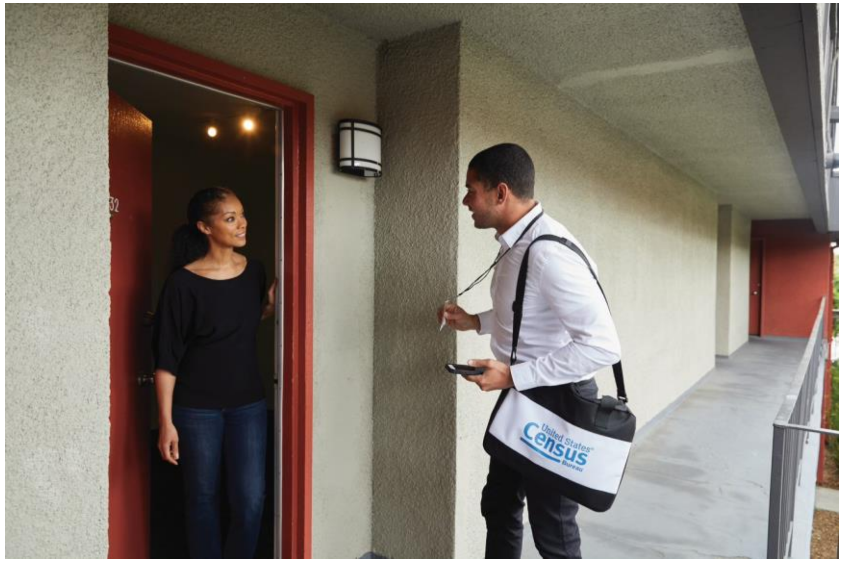
AUBER Webinar
January 21, 2020

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

Our Commitment to Data Stewardship

Data stewardship is central to the Census Bureau's mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



Shape your future START HERE >

United States® Census 2020

Upholding our Promise: Today and Tomorrow

We cannot merely consider privacy threats that exist today.

We must ensure that our disclosure avoidance methods are also sufficient to protect against the threats of tomorrow!

The Privacy Challenge

Every time you release any statistic calculated from a confidential data source you "leak" a small amount of private information.

If you release too many statistics, too accurately, you will eventually reveal the entire underlying confidential data source.

The Census Bureau's Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.

Stopped publishing small area data	Whole-table suppression	Data swapping	Formal Privacy
1930	1970	1990	2020

The Growing Privacy Threat

More Data and Faster Computers!

In today's digital age, there has been a proliferation of databases that could potentially be used to attempt to undermine the privacy protections of our statistical data products.

Similarly, today's computers are able to perform complex, large-scale calculations with increasing ease.

These parallel trends represent new threats to our ability to safeguard respondents' data.

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 150 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: 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…

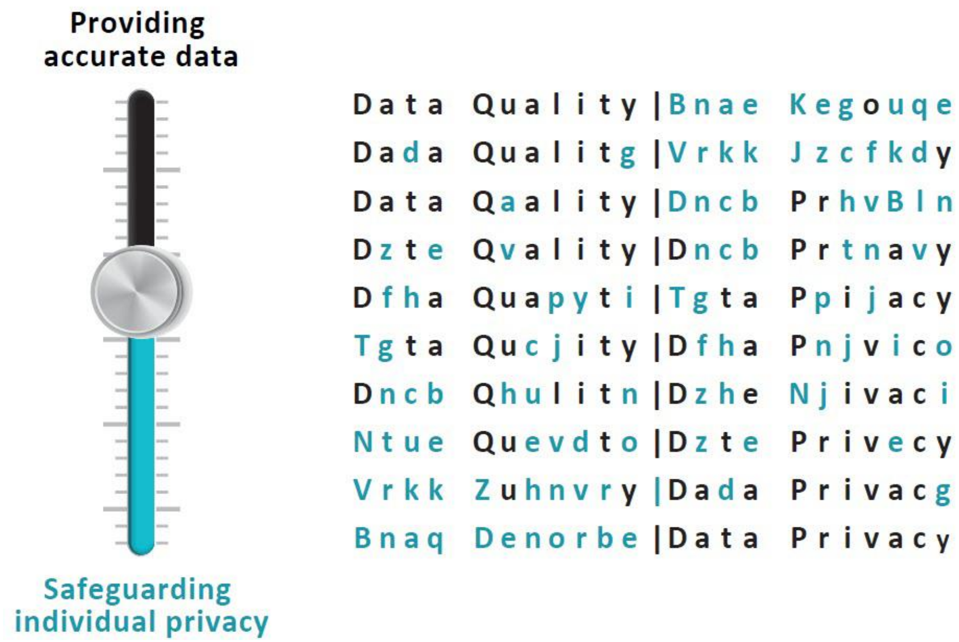
-now, or at any point in the future!

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.



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."

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

ε

Epsilon

Shape your future START HERE >

United States® Census 2020

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.

Shape your future START HERE >

United States Census 2020

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.
- Uses Privacy-Loss Budget of ε=6 (ε=4 for person records, ε=2 for household records)

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

Send feedback (by Summer 2020) to:
dcmd.2010.demonstration.data.products@census.gov

Known Issues – Work is Ongoing!

- **There are two sources of error in the TopDown Algorithm (TDA):**
  - Measurement error due to differential privacy noise
  - Post-processing error due to statistical inference creating non-negative integer counts from the noisy measurements
- **Post-processing error tends to be much larger than differential privacy error**
- **Positive bias in small counts/negative bias in large counts is the result of**
  - Invariants
  - Post-processing error specifically introduced by our L2 optimization routine
- **Improving post-processing is not constrained by differential privacy**
- **Techniques to improve post-processing error may be drawn from demography, statistics, computer science, operations research, econometrics, etc. without increasing the privacy-loss budget**

Questions?

Disclosure Avoidance and the 2020 Census Website
https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

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