Balancing Respondent Privacy and Reliability in the 2020 Census and Beyond: Implications for Data Users

Constance F. Citro, Senior Scholar, CNSTAT

APDU Annual Conference
July 9, 2019
In response to concerns around protecting the privacy of census respondents, the Census Bureau is making changes to their data products starting with the 2020 Decennial Census tables.

They are also considering changes to the American Community Survey products and others.

This panel will discuss the tradeoffs between data utility and respondent privacy, while clarifying the “differential privacy” approach the Bureau plans to use for 2020.
Explaining “Differential Privacy” in Lay Terms (oxymoron?)

- What is the problem DP will try to solve for 2020?
  - The Census Bureau has conducted research with 2010 summary tables that demonstrates a risk of potential re-identification of a significant percentage of individual respondents—“potential” because an attacker can’t be 100% confident of a re-identification.

- The Census Bureau therefore plans for 2020 data products to replace traditional protection methods (swapping, topcoding, etc.) with a new privacy protection concept that mathematically produces a given distortion (from injecting noise) into data elements for a specified privacy “budget” ($\epsilon$)—the larger the value of epsilon, the less privacy protection and the more accurate the data.
Implications for Users

- The Census Bureau will not release as much data for 2020 as it did in 2010.
- The ability to impose a privacy budget constraint and determine the amount of noise needed requires *pre-specification* of all tables.
- The pressing need, therefore, is to prioritize users’ data needs (“use cases”).
- A key problem is that users’ needs for detailed race/ethnicity/origins and family/person characteristics make it difficult to protect privacy and even to determine a feasible algorithm for implementing a specified epsilon.
Don’t Worry (too much) about the ACS, Yet

➢ The mathematics for DP have not been worked out for data sets based on complex survey designs, such as the ACS.
➢ However, the Census Bureau will likely reduce the data that it releases publicly for the ACS.
➢ For both 2020 and the ACS, PUMS data are at risk.
➢ It is imperative that the Census Bureau and users work together to determine data needs—content and acceptable levels of noise.
➢ After the panelists offer their perspectives, I will encourage Q&A and input on use cases.