The New York Times https://nyti.ms/2E4UeZQ

To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

Guaranteeing people's confidentiality has become more of a challenge, but some scholars worry that the new system will impede research.

By Mark Hansen

Dec. 5, 2018

When the Census Bureau gathered data in 2010, it made two promises. The form would be "quick and easy," it said. And "your answers are protected by law."

But mathematical breakthroughs, easy access to more powerful computing, and widespread availability of large and varied public data sets have made the bureau reconsider whether the protection it offers Americans is strong enough. To preserve confidentiality, the bureau's directors have determined they need to adopt a "formal privacy" approach, one that adds uncertainty to census data before it is published and achieves privacy assurances that are provable mathematically.

The census has always added some uncertainty to its data, but a key innovation of this new framework, known as "differential privacy," is a numerical value describing how much privacy loss a person will experience. It determines the amount of randomness — "noise" — that needs to be added to a data set before it is released, and sets up a balancing act between accuracy and privacy. Too much noise would mean the data would not be accurate enough to be useful — in redistricting, in enforcing the Voting Rights Act or in conducting academic research. But too little, and someone's personal data could be revealed.

On Thursday, the bureau will announce the trade-off it has chosen for data publications from the 2018 End-to-End Census Test it conducted in Rhode Island, the only dress rehearsal before the actual census in 2020. The bureau has decided to enforce stronger privacy protections than companies like Apple or Google had when they each first took up differential privacy.

Cynthia Dwork, the Gordon McKay professor of computer science at Harvard and one of the inventors of differential privacy, says that it is "tailored to the statistical analysis of large data sets" — precisely the issue facing the census with its mandate from Title 13 of the U.S. Code to keep each person's information private, and its responsibility to provide useful data.

At the root of the problem are the tables of aggregate statistics that the bureau publishes. There are hundreds of tables — sex by age, say, or ethnicity by race — summarizing the population at several levels of geography, from areas the size of a city block all the way up to the level of a state or the nation. In 2010, the bureau released tables with nearly eight billion numbers in all. That was about 25 numbers for each person living in the United States, even though Americans were asked only 10 questions about themselves. In other words, the tables were generated in so many ways that the Census Bureau ended up releasing more data in aggregate then it had collected in the first place.

In 2003, Irit Dinur and Kobbi Nissim, then with the NEC Research Institute, warned about extensive data publication and the inadequacy of many existing attempts to ensure confidentiality, including those used by the Census Bureau in 2000 and 2010. The two researchers' "database reconstruction theorem" provides a road map for someone to turn collections of summary tables into approximate records on individuals.

An activist during a news conference in New York in April. A multistate lawsuit was announced to block the Trump administration from adding a question about citizenship to the census. Drew Angerer/Getty Images

For the census, this is particularly worrisome, especially if a question about citizenship is added to the 2020 census, as the Trump administration has proposed. "I think it is crystal clear what the potential harm is from poorly protected tabular summaries," said John Abowd, associate director for research and methodology at the Census Bureau, who became an early proponent of differential privacy.

In November 2016, the bureau staged something of an attack on itself. Using only the summary tables with their eight billion numbers, Mr. Abowd formed a small team to try to generate a record for every American that would show the block where he or she lived, as well as his or her sex, age, race and ethnicity — a "reconstruction" of the person-level data.

Each statistic in a summary table leaks a little information, offering clues about, or rather constraints on, what respondents' answers to the census *could* look like. Combining statistics from different aggregate tables at different levels of geography, we start to get a picture of the demographics of who is living where.

On the face of it, finding a reconstruction that satisfies all of the constraints from all the tables the bureau produces seems impossible. But Mr. Abowd says the problem gets easier when you notice that these tables are full of zeros. Each zero indicates a combination of variables — values for one or more of block, sex, age, race and ethnicity — for which no one exists

in the census. We might find, for example, that there is no one below voting age living on a particular block. We can then ignore any reconstructions that include people under 18 living there. This greatly reduces the set of viable reconstructions and makes the problem solvable with off-the-shelf software.

As an illustration, following the details available in public presentations from Mr. Abowd and his colleagues, we were able to perform our own reconstruction experiment on Manhattan. Roughly 1.6 million people are divided among 3,950 census blocks — which typically correspond to actual city blocks. The summary tables we needed came from the census website; we used simple tools like R and the Gurobi Optimizer; and within a week we had our first results.

By this summer, Mr. Abowd and his team had completed their reconstruction for nearly every part of the country. When they matched their reconstructed data to the actual, confidential records — again comparing just block, sex, age, race and ethnicity — they found about 50 percent of people matched exactly. And for over 90 percent there was at most one mistake, typically a person's age being missed by one or two years. (At smaller levels of geography, the census reports age in five-year buckets.)

This level of accuracy was alarming. Mr. Abowd and his peers say that their reconstruction, while still preliminary, is not a violation of Title 13. Instead it is seen as a red flag that their current disclosure limitation system is out of date.

The bureau has long had procedures to protect respondents' confidentiality. For example, census data from 2010 showed that a single Asian couple — a 63-year-old man and a 58-year-old woman — lived on Liberty Island, at the base of the Statue of Liberty.

That was news to David Luchsinger, who had taken the job as the superintendent for the national monument the year before. On Census Day in 2010, Mr. Luchsinger was 59, and his wife, Debra, was 49. In an interview, they said they had identified as white on the questionnaire, and they were the island's real occupants.

Before releasing its data, the Census Bureau had "swapped" the Luchsingers with another household living in another part of the state, who matched them on some key questions. This mechanism preserved their privacy, and kept summaries like the voting age population of the island correct, but also introduced some uncertainty into the data.

David Luchsinger and his wife, Debbie, out for a walk in their neighborhood, Liberty Island, in 2011, or one year after the 2010 census. Mustafah Abdulaziz for The New York Times

The bureau's attack on itself showed that swapping wasn't enough. Swapping focused on people who were isolated like the Luchsingers or who had characteristics that made them stand out in their neighborhood — the cells in the tables with only a single person. But as we have seen, it is the zeros in the tables that give us power to perform the reconstruction in the first place. To guarantee privacy, we need to consider the zeros as well.

On Thursday, the Census Bureau will reveal the details of applying differential privacy to its 2018 End-to-End Census Test, how it will control the level of noise in the summary tables to guarantee privacy. To make its choice, the bureau experimented with different values; used each to generate sets of summary tables from the 2010 census; and compared their accuracy.

Mr. Abowd wrote via email that the bureau would release a large batch of the tables from its experiments, along with the production code used to generate them: "Users of all stripes will be able to use those simulation results to assess average error for their own use cases at all levels of geography." In presentation materials for Thursday's announcement, special attention is paid to lessening any problems with redistricting: the potential complications of using noisy counts of voting-age people to draw district lines. (By contrast, in 2000 and 2010 the swapping mechanism produced exact counts of potential voters down to the block level.)

The Census Bureau has been an early adopter of differential privacy. Still, instituting the framework on such a large scale is not an easy task, and even some of the big technology firms have had difficulties. For example, shortly after Apple's announcement in 2016 that it would use differential privacy for data collected from its macOS and iOS operating systems, it was revealed that the actual privacy loss of their systems was much higher than advertised.

Some scholars question the bureau's abandonment of techniques like swapping in favor of differential privacy. Steven Ruggles, Regents Professor of history and population studies at the University of Minnesota, has relied on census data for decades. Through the Integrated Public Use Microdata Series, he and his team have regularized census data dating to 1850, providing consistency between questionnaires as the forms have changed, and enabling researchers to analyze data across years.

"All of the sudden, Title 13 gets equated with differential privacy — it's not," he said, adding that if you make a guess about someone's identity from looking at census data, you are probably wrong. "That has been regarded in the past as protection of privacy. They want to make it so that you can't even guess."

"There is a trade-off between usability and risk," he added. "I am concerned they may go far too far on privileging an absolutist standard of risk."

In a working paper published Friday, he said that with the number of private services offering personal data, a prospective hacker would have little incentive to turn to public data such as the census "in an attempt to uncover uncertain, imprecise and outdated information about a particular individual."

Timothy Donald Jones, a Ph.D. candidate in statistics at Columbia University, contributed reporting.

Mark Hansen is a professor of journalism at Columbia University, where he also serves as the director of the David and Helen Gurley Brown Institute for Media Innovation. You can follow him on Twitter at @cocteau.