Society | DOI:10.1145/3329719 Logan Kugler

Protecting the 2020 Census

A new framework is being used to secure the 2020 U.S. Census from database reconstruction attacks.

In 2020, the people of the U.S. will stand up and be counted, according to the provisions in the U.S. Constitution that stipulate a census may take place every decade. It's a tradition dating back to 1790, when the first national census was conducted.

This tradition is turning to a newer technique to stay secure in the 21st century.

Back in 2003, researchers Irit Dinur and Kobbi Nissim of the NEC Research Institute published a paper explaining how they had identified theoretical vulnerabilities in the summary data published with confidential databases. In some cases, the researchers found, the summary data—a high-level picture of the data from individual records in a database—could be used to reconstruct the private database. That meant attackers could use the public summary of the data to reconstruct what people had disclosed privately.

On paper, these types of database reconstruction attacks presented a possible threat to confidential databases that published summary data. The U.S. Census is a prime example of such a database.

For a long time, the paper remained a warning about a theoretical threat; until the last decade, when a dramatic increase in both computer speed and the efficiency of NP-hard problem solvers turned the theoretical threat into a practical peril, according to research published by U.S. Census Bureau employees.

One of those employees, John Abowd, associate director for research and methodology at the Bureau, worked with a team to investigate whether advances in computing power could enable database reconstruction attacks on the U.S. Census.

The results were shocking.

Abowd and his team retroactively used database reconstruction techniques on these public data summaries, and found they could use advanced computational power and techniques to recreate private data that was never meant to be public.

IMAGE BY MARIA DRYFHOUT

In fact, Abowd and his team found they could reconstruct all the records contained in the database with approximately 50% accuracy. When they allowed a small error in the age of an individual, the accuracy with which they could associate public data with individuals went up to 70%. And if they allowed getting *one* piece of personal information like race or age wrong, but everything else right, their reconstruction was more than 90% accurate.

"The vulnerability is not a theoretical one; it's an actual issue. The systems being used [for the census] were vulnerable," says Abowd.

The solution, it turns out, was just as modern as the problem.

A Modern Solution to a Modern Problem

By law, the U.S. Census Bureau is prohibited from identifying "the data furnished by any particular establishment or individual." That is why the Census Bureau publishes summary data, or a high-level view of the sex, age, race, and other household details of Americans by state.

The main data product that comes out of the Census is Summary File 1, which constitutes the "main dissemination of census results," says Abowd. Summary File 1 contains a lot of data that demographers use, like age, race, and ethnicity segmented by gender, as well as household composition statistics.

According to the Census Bureau, Summary File 1 "includes population and housing characteristics for the total population, population totals

for an extensive list of race ... and Hispanic or Latino groups, and population and housing characteristics for a limited list of race and Hispanic or Latino groups."

Abowd and his team took Summary File 1 data from the 2010 Census and subjected it to a database reconstruction attack, and found they were able to uncover privately disclosed data with some accuracy.

When Abowd presented his findings to senior executives at the Census Bureau, the agency interpreted the ability to reconstruct private data as a breach of the confidentiality obligation it had under law. In that context, it was decided that action needed to be taken to correct the vulnerability before the 2020 Census.

The Bureau's executive team discussed the issue, then made the decision to use a statistical method called differential privacy to secure the Census process.

Explains Jonathan Ullman, an assistant professor of computer science at Northeastern University with specialties in cryptography and privacy, differential privacy is a way to prevent attackers from reconstructing databases by adding statistical "noise" to those databases.

The Bureau's executive team made the decision to use a statistical method called differential privacy to secure the Census process.

Statistical noise refers to altering the aggregate results that come from a database like the Census, so it is more difficult to use these aggregate results to identify the original data collected. Ullman offers an example: rather than reporting the median income of a resident of a town in the U.S. as $66,500, you could choose a random number between $66,000 and $67,000 to add noise.

"Adding this noise makes it harder for someone to reconstruct the database or otherwise breach privacy by combining many statistics," says Ullman.

Ideally, the amount of noise should be pretty small, so the statistics can still be used by researchers, thousands of whom rely on Census data for their work. After all, statisticians and researchers are "already used to thinking of their data as containing various sources of error," such as sampling error and response bias, according to Ullman.

However, Ullman cautions, "We have to be careful about how much noise we add and how we do it," so the data strikes the right balance between confidential and useful. Adding the right amount of noise can make it more difficult to reconstruct the data-

ACM News

A Healthy Dose of Wearables

The idea of carrying a small device to monitor your health is not new. From the 1960s television show "Star Trek," which featured a fictional handheld scanning device called the Tricorder, to more recent smartwatches and wearable devices, portable technology that aims to improve medicine and healthcare has advanced steadily.

Wearables are reshaping medicine in significant ways. Last September, for example, Apple received approval from the U.S. Food and Drug Administration (FDA) to include medical-grade heart monitoring in its Series 4 Apple Watch. In addition, the watch detects falls, and can alert emergency responders if such an event takes place.

Manufacturers are also introducing glucose and blood pressure monitoring, breast cancer detection, and more advanced sleep monitoring and feedback into wearables. Embedded sensors, along with wireless technology and the Internet of Things, are pushing the boundaries of medicine into new frontiers.

Smartwatches and wearables can spot potential medical problems, improve patient behavior, and boost compliance. The Apple Watch is perhaps the highest-profile smart wearable device, but other manufacturers are now streaming into the market with wearables that address an array of health challenges.

For example, iSono Health has introduced a three-dimensional (3D) ultrasound system that uses a bra to detect unusual lumps and masses in breasts, and transmits data to a smartphone or tablet.

Medical device maker Omron has received FDA approval on a blood pressure monitor that looks like a smartwatch and connects to a smartphone.

"These devices change healthcare in significant ways," says Arielle Trzcinski, a senior analyst at Forrester Research. "An individual gains greater insight into what is happening inside his or her body without having to visit a doctor and sit in an exam room."

Connected wearables offer other advantages. One of the most significant is access to data not previously available—or feasible to collect. For instance, the Apple Heart Study, conducted by Stanford Medicine, now has more than 400,000 participants, making it the largest screening study ever conducted for atrial fibrillation.

Wearable medical devices and the data they collect may lead to different treatment approaches, and introduce insurance rates at least partially based on patient incentives. "These devices and technology in general can aid in patient care and help alleviate burnout for clinicians. But these systems must be designed so that clinicians don't spend hours reviewing meaningless data," says Adrienne Boissy, a neurologist and chief experience officer for Cleveland Clinic Healthcare systems.

Not surprisingly, the quality of data and whether it can be used to accurately identify and diagnose specific issues is critically important. Unlike fitness devices, medical devices must be precise. But security and privacy concerns also exist, including who controls and owns data as it streams across devices, systems, and companies. Finally, says Trzcinski, organizations must reexamine "the dynamics of how clinicians and patients interact" and how billing and reimbursements take place.

Nevertheless, the ability to integrate these devices into people's lives could change thinking, behavior, and actions—and lead to healthier, happier people.

Concludes Trzcinski, "We're going to continue to see remarkable innovation. We're going to see devices that are smaller and smarter. As the technology improves and underlying algorithms become more precise and accurate through machine learning, the chance to actually improve people's lives will grow."

—Samuel Greengard is an author and journalist based in West Linn, OR, USA.

base, while also leaving the data sufficiently useful for researchers.

Differential privacy can make sure you're drawing the right balance between noise in your data and the usefulness of your data. Researchers Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith presented a paper at the 2006 Theory of Cryptography Conference, "Calibrating Noise to Sensitivity in Private Data Analysis," showing how to set up a mathematical system that allows parametric control over a risk that can be quantified, while formalizing the amount of noise needed to be added to protect the data and proposing a generalized mechanism for doing so.

"It was specifically designed to provide mathematical assurances that you had controlled the risk of database reconstruction, specifically that you controlled the potential harm from re-identification caused by an attacker building too accurate an external image of your data," says Abowd.

This is why differential privacy was picked by the Census Bureau to defend its data.

"It's a mathematical framework for understanding what 'ensuring privacy' means," says Ullman. "The framework was specifically tailored to understanding how to protect privacy in statistical analysis of large datasets, which is exactly the problem the Census faces."

Abowd began experimenting with differential privacy frameworks in 2008 as part of other work for the Census Bureau, which produces a number of data products aside from the Census itself. However, it wasn't until 2016, after he conducted a database reconstruction attack on past Census data, that the need to use differential privacy on all Census data became apparent.

Census Bureau management agreed with Abowd that differential privacy was the solution to the problem, so Abowd and a team of computer scientists and engineers got to work implementing it.

Balancing Privacy with Usability

Abowd put together a team of computer scientists and engineers in short order to combat the threat. The team includes science lead Dan Kifer, a professor of computer science at Penn State University, and engineering lead Simson Garfinkel, previously a computer scientist at the National Institute of Standards and Technology (NIST). The team is currently working to apply differential privacy to the Census' upcoming efforts for 2020.

"The framework was specifically tailored to understanding how to protect privacy in statistical analysis of large databases, which is exactly the problem the Census faces."

It is not an easy task.

"We have to do it fast, and we have to do it well," says Abowd. Though he readily admits the tight timeline and volume of work are heavy burdens, and these are not the only obstacles.

The community of researchers who use Census data will be dealing with data in 2020 that has a new system of protection applied to it, and not everyone is happy about that.

One outspoken critic is Steven Ruggles, Regents Professor of History and Population Studies at the University of Minnesota, and director of the Institute for Social Research and Data Innovation, which is focused on advancing "our knowledge of societies and populations across time and space, including economic and demographic behavior, health, well-being, and human-environment interactions." Ruggles regularly uses Census data in his work, and says the use of differential privacy could limit the ability of researchers to find useful insights in that data.

"The fundamental problem is loss of accuracy of the data," says Ruggles. "In the case of tabular small-area data, noise injection will blur the results, potentially leading investigators and planners to miss patterns in the data. For example, the noise injection could lead to underestimation of residential segregation."

Ruggles also does not believe the implementation of differential privacy on U.S. Census data is even necessary. "There has never been a documented case of anyone's identity being revealed in a public-use data product, so it is a huge overreaction."

Ullman, on the other hand, sees differential privacy as the best solution available to prevent database reconstruction attacks, while still keeping the data of the Census usable.

Because the Census has an enormous dataset, Ullman says it is possible to release huge quantities of summary statistics with manageable amounts of noise. Differential privacy then quantifies how releasing additional summary statistics will increase privacy risks, making it possible to "weigh the harm to privacy against the public benefits in a sensible way."

"There is simply no competing framework right now that has the potential to offer all of these benefits," Ullman says. C

Further Reading

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006) Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S., Rabin T. (eds) Theory of Cryptography. TCC 2006. *Lecture Notes in Computer Science*, vol 3876. Springer, Berlin, Heidelberg http://bit.ly/2DbERfW

Hansen, M.
To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data, *The New York Times*, Dec. 5, 2018
https://nyti.ms/2UITL4n

Garfinkel, S., Abowd, J., and Martindale, C.
Understanding Database Reconstruction Attacks on Public Data, *ACM Queue*, Nov. 28, 2018
https://queue.acm.org/detail.cfm?id=3295691

Logan Kugler is a freelance technology writer based in Tampa, FL, USA. He has written for over 60 major publications.