New tools safeguard Census data about where you live and work

May 18 2017, by Robin A. Smith

In October 2012, as Hurricane Sandy bore down on the densely populated U.S. East Coast, the state of New Jersey needed information fast. State planners and emergency managers turned to U.S. Census Bureau data about the people living and working in the affected area to identify the communities that would be hardest hit, and come up with a plan for recovery in the months that followed.

A team led by Duke University, in collaboration with the Census Bureau, has developed new methods that enable people to learn as much as possible from Census data and other government workforce statistics for things like disaster management, policy-making and funding decisions, while guaranteeing that no one can trace the data back to your household or business.

Census-related statistics are used to allocate more than 400 billion dollars annually for disaster relief, job training centers, roads and other services. At the same time, Americans entrust the Census Bureau and other agencies with their personal information on the understanding that their answers will be kept confidential.

Duke assistant professor of computer science Ashwin Machanavajjhala and graduate student Samuel Haney described their approach on May 18 at the Association for Computing Machinery's 2017 SIGMOD/PODS Conference in Chicago.

Any time you fill out a Census Bureau survey or census, your answers

are combined with others to produce summary statistics about the local population in each city and rural area: how many people live and work there, what jobs they do, how they commute to work and other characteristics.

These de-identified data are released in aggregated form and mined for patterns and insights that have an enormous impact on your life, from whether a business decides to relocate or expand to your area, to how many police officers and firefighters your town has and where to build new hospitals and schools.

The Census Bureau uses a variety of measures to ensure that no one can reverse-engineer the data and identify individuals within them. The challenge, Machanavajjhala said, is to enable third parties to sift through the data to make discoveries about the respondents as a group, while revealing as little as possible about any individual or business in it.

One approach to this balancing act, first proposed in 2006, involves a set of techniques called "differential privacy."

With differential privacy, a person can share their personal information without worrying that someone analyzing the aggregated data might be able to figure out which information is hers. No one can identify your data, even if they have other information about you.

Although differential privacy was introduced more than 10 years ago, it is just now beginning to be put to use more widely for collecting and sharing sensitive data. Apple has implemented differential privacy techniques in iOS 10, the latest mobile operating system for the iPhone. Google has done the same for its Chrome Web browser.

"Today, differential privacy is considered a gold standard for analyzing sensitive data," Machanavajjhala said.

Differential privacy techniques typically work by computing the true answer and then adding random noise to the output. The goal, Machanavajjhala said, is to ensure that adding or removing any single record or individual from the database doesn't significantly affect the outcome. But critics of differential privacy argue that achieving that goal requires introducing too much noise to glean accurate insights.

That goal is based on a mathematical definition of privacy that may be overly strict for certain applications, Machanavajjhala said. "And in many cases, it may not match up with what is required by law, or what users think is meant by privacy."

Instead, the researchers tried a new approach. They took the privacy protections required by law, and adapted them into a customized definition of privacy similar to differential privacy. Then they developed algorithms that injected just enough noise to satisfy that definition and uphold the law.

The researchers ran an experiment where they applied their algorithms to real-world employment data underlying an online mapping tool produced by the U.S. Census Bureau, called OnTheMap for Emergency Management.

In the wake of Hurricane Sandy, the mapping tool has provided accurate estimates of the number of affected workers by race, ethnicity, industry or other characteristics, to figure out which groups of people or types of businesses were most in need of aid.

Users could also use the data set to view where businesses are located and how employees travel to work to determine the impact of the hurricane on commuters.

But in their experiment, the researchers were able to prove,

mathematically, that such answers wouldn't get someone any closer to inferring information about any single person or business that might violate privacy regulations—such as whether an employee held a job at a particular workplace, or precisely what fraction of a company's workforce belonged to a certain race or had a certain level of education.

Their study showed it is possible to release such data to the public for analysis and guarantee it safe against unwanted leaks, with comparable or even better accuracy of the results than current methods—which don't make similar privacy guarantees.

The algorithms aren't specific to U.S. Census Bureau data. The techniques they developed are applicable to other employment-related statistics produced by other countries or agencies as well, such the U.S. Bureau of Labor Statistics and the U.S. Bureau of Economic Analysis.

More information: "Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics," Samuel Haney, Ashwin Machanavajjhala, John Abowd, Matthew Graham, Mark Kutzbach and Lars Vilhuber. ACM SIGMOD/PODS Conference, Chicago, IL, May 14-19, 2017. DOI: 10.1145/3035918.3035940

Provided by Duke University