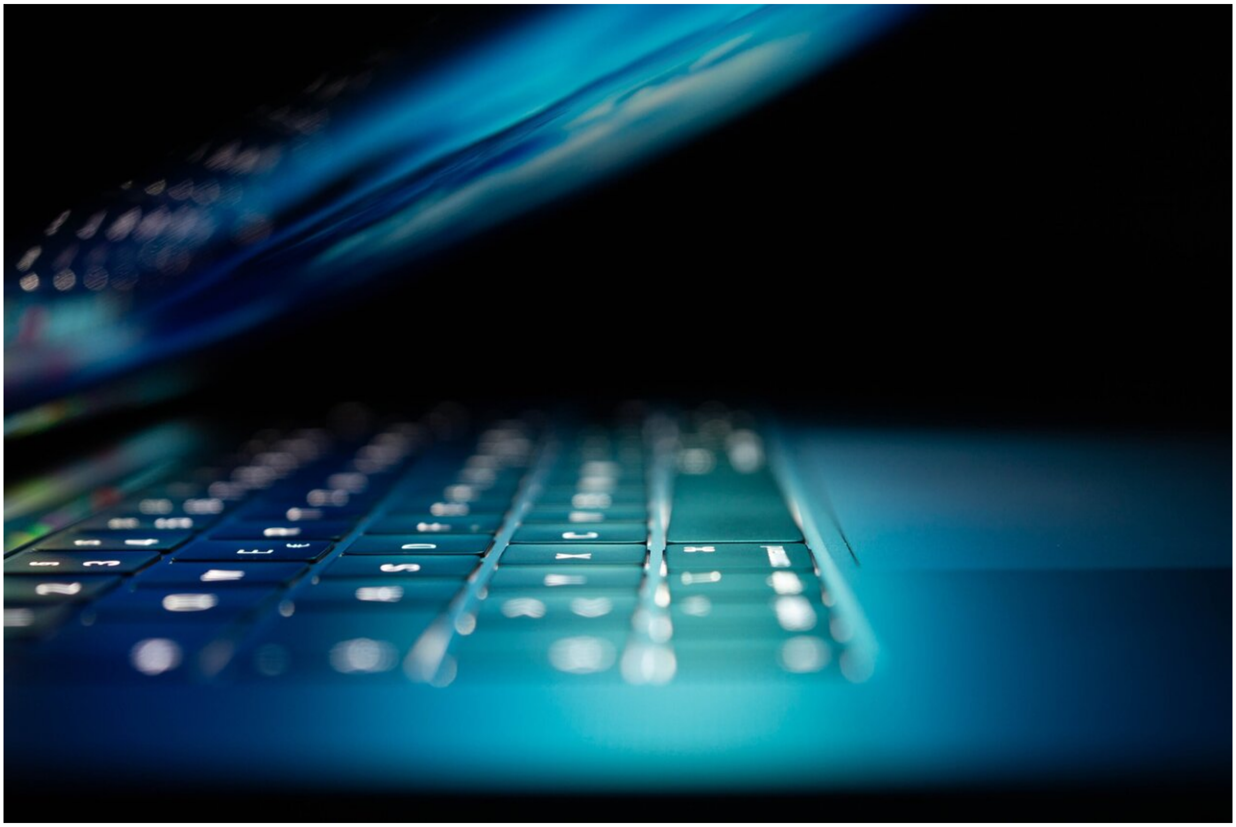
TechXplore

Differential privacy the correct choice for the 2020 US Census

May 18 2022



Credit: Unsplash/CC0 Public Domain

The U.S. Census Bureau has long struggled to balance the accuracy and privacy of its decennial census data. High-impact use cases such as funding allocation and redistricting make the accuracy of this data especially crucial. On the other hand, census data privacy is not only required by law, but is also important for protecting vulnerable populations and ensuring a high response rate. Due to serious privacy concerns about its previous de-identification method, swapping, the Census Bureau recently switched to a newer method: differential privacy.

Differential privacy (DP), put simply, is a mathematical concept that keeps people's personal information private by injecting "noise"—small, random changes—into the data. Concerns have been raised that this noise will artificially deflate the reported populations of minority groups, resulting in lost funding. A team of researchers from Columbia Engineering's department of computer science decided to study these claims, asking: is this risk specific to DP? They conducted both theoretical and empirical analyses comparing DP algorithms to swapping algorithms.

New study supports switch to differential privacy

Their findings, which will be presented May 23, 2022, at the IEEE Symposium on Security and Privacy in San Francisco, support the Census Bureau's switch to differential privacy as a de-identification mechanism for the 2020 Census, and show that swapping produces poor accuracy for minority groups. Furthermore, swapping places a disproportionate privacy burden on minority groups while DP provides a stronger privacy guarantee.

"The more that we can understand about the impact of disclosure avoidance procedures on data, the better," said danah boyd, Partner Researcher at Microsoft Research and founder of Data & Society, who was not involved in the study. "This new study reveals important insights about how different mechanisms impact vulnerable communities in different ways. Our country depends on data to allocate resources and representation. The stakes are high. This paper helps us see the technical challenges of producing high-quality data without risking people's privacy,"

Study originated from a class on anonymity and privacy

The study grew out of a project for a spring '21 class on anonymity and privacy taught by Professors Steven Bellovin, co-author of the paper, and Alex Abdo of Columbia's Knight Institute. "We had a number of interesting projects in this class and this one really stood out to me," said Bellovin, the Percy K. and Vida L.W. Hudson Professor of Computer Science and an affiliate faculty at Columbia Law School. "To our knowledge, we are the first to directly compare the effects of swapping to the effects of DP on minority under-representation across a wide range of parameter settings."

Miranda Christ, a Ph.D. student, and Sarah Radway, a senior in the class, were initially concerned by articles such as the New York Times opinion piece discussing census data inaccuracy due to the noise added by DP. When Sarah and Miranda realized that census data has included noise for many years, due to previous disclosure avoidance methods such as swapping, they were surprised to find a lack of research comparing the relative inaccuracy of swapping and DP. In their project, they aimed to determine how the two privacy methods compared, in terms of both accuracy and privacy. With Bellovin's encouragement, they decided to go beyond the classroom project, consulting experts in the field such as Rachel Cummings, assistant professor of industrial engineering and operations research at Columbia Engineering,

"It's always great when academic research can touch the real world," said Radway, now a Ph.D. student with Susan Landau, professor of cybersecurity and policy at Tufts University. "This is a serious controversy—there was even a lawsuit in a Federal court. We showed that the Census Bureau's judgments were correct, and that they made the right decision."

Christ, a Ph.D. student at Columbia co-advised by Computer Science Professors Tal Malkin and Mihalis Yannakakis, added, "This research is especially important now, as the Census Bureau begins to modernize its disclosure avoidance methods for its other surveys, such as the American Community Survey. It will also help inform similar decisions in other similar settings."

Swapped data more inaccurate for minorities—unlike differential privacy

The researchers demonstrated that the inaccuracy added by swapping is more harmful than that of DP. In particular, they showed that when swapping is implemented with sufficient privacy, its accuracy is no better than, and often much worse than, that of differential privacy. Swapped data is more inaccurate for more diverse counties, and even more inaccurate for minorities—this is not the case for differential privacy. The study also shows that minority groups are at a higher risk of identification in swapped data.

More information: Differential privacy and swapping: Examining de-identication's impact on minority representation and privacy preservation in the U.S. Census.", Conference: IEEE Symposium on Security and Privacy, DOI: 10.1109/SP46214.2022.00135

Provided by Columbia University School of Engineering and Applied

Science

Citation: Differential privacy the correct choice for the 2020 US Census (2022, May 18) retrieved 9 September 2024 from https://techxplore.com/news/2022-05-differential-privacy-choice-census.html