DATA & SOCIETY | The Leadership Conference | CENTER ON POVERTY and INEQUALITY GEORGETOWN LAW | ECONOMIC SECURITY and OPPORTUNITY INITIATIVE

FACT SHEET

Differential Privacy in the 2020 Census

NEW CONFIDENTIALITY PROTECTIONS & THE IMPLICATIONS FOR DATA USERS

JAE JUNE LEE & CARA BRUMFIELD

KEY TAKEAWAYS

- The Census Bureau is modernizing how it protects the confidentiality of census responses. Confidentiality protections for the 2020 Census relies on "differential privacy" (or "formal privacy").
- Differential privacy allows the bureau to provide robust and measurable confidentiality protections against the evolving challenges presented by advances in computer science and the growing availability of personal information online and from commercial providers.
- Like all disclosure avoidance methods, differential privacy involves a trade-off between confidentiality and the utility of published statistics. The more statistics that are published, and the closer those statistics match the underlying, confidential data, the greater the risk to the confidentiality of individual respondents.
- Researchers, civil rights groups, and other data users should actively engage in Census Bureau consultations about which data products and statistics to publish.

The Census Bureau is a steward of the data it collects: the agency is tasked with deciding how to best produce meaningful statistics while ensuring that the statistics do not disclose, or allow others to discover, confidential information about individual respondents. The Census Bureau has a legal obligation to protect confidentiality and recognizes that unacceptable levels of confidentiality loss may undermine the public's trust in the bureau and people's willingness to participate in future censuses. Starting with the 2020 Census, the bureau will modernize its protections by adopting a mathematical definition of confidentiality called "differential privacy." Differential privacy offers notable benefits, including robust and measurable guarantees of confidentiality.

Like all disclosure avoidance methods, it involves a trade-off between confidentiality and the utility of publicly available data. Users of census data have voiced concerns that the new measures may limit the utility of census statistical products and public-use microdata—impacting an array of uses, including research, local policy decisions, and the enforcement of civil rights laws. Yet, civil society organizations and other stakeholders are also deeply aware of the serious confidentiality concerns among many of the communities they advocate for and represent. Users of census data should actively engage with the Census Bureau in decisions about how to manage the tradeoff between confidentiality and the usefulness and availability of census data.

THE BUREAU IS MODERNIZING ITS CONFIDENTIALITY PROTECTIONS

In making statistics available to the public, the bureau faces a fundamental challenge: the more statistics that are published, and the closer those statistics match the underlying, confidential data, the greater the probability that a bad actor can reconstruct the underlying data and identify individual respondents. Historically, the bureau employed a collection of methods to prevent disclosure of confidential information. This included inserting random error (or "noise") into statistics on selected populations and households, "swapping" specific data elements, and limiting the amount of published information. While these methods deliberately introduce error into published statistics, disclosure avoidance is not the only source of error in the decennial census: for example, nonsampling errors, such as measurement and nonresponse errors (i.e., undercounts), contribute to a lack of precision in statistics.

The bureau's internal research, however, has led the agency to believe that disclosure avoidance methods used in past censuses will not be sufficient for the 2020 Census and beyond. Advances in computer science and statistical techniques, along with the growing abundance of personal information online and from commercial providers, present new and evolving threats to confidentiality. In response, the bureau has turned to differential privacy to provide robust and measurable guarantees of confidentiality. While prior confidentiality protections

were applied to specified elements of the data, the bureau will now introduce a **controlled quantity of noise** to every statistic it produces. This approach allows the bureau to quantify how accurate the published statistics are and how much confidentiality is lost—and to adjust the noise (controlled by the variable "epsilon") to change the balance of utility and confidentiality. Put another way, published statistics are approximate—and not exact matches—to the underlying, confidential data, and this proximity can be measured and calibrated.

USER FEEDBACK WILL INFORM HOW DIFFERENTIAL PRIVACY IS IMPLEMENTED

The new disclosure avoidance system requires the bureau to decide in advance which **statistical products it will publish** and the geographic levels (e.g., state, block-group) of the data it will provide. As these decisions are made, the bureau must consider the myriad uses of these data and the implications for communities. Although the bureau faces tight timelines as the 2020 Census fast approaches, it must make time to communicate its decisions and listen to stakeholder input.

For example, the National Congress of American Indians (NCAI) shared with the bureau **concerns** that the implementation of differential privacy could introduce substantial amounts of noise into statistics for small populations living in remote areas, potentially diminishing the quality of statistics about tribal nations. As a result, the bureau arranged formal consultations with tribal nations to better understand how the data are used in local decision-making, research, and the distribution of federal funds.

The bureau is still determining how it will apply brand-new methodologies to produce public-use microdata and some of the decennial census tables, particularly for ones that combine household and individual characteristics such as the detailed race and ethnicity tables. However, the Census Bureau has provided outlines for some of the data products that it is currently able to produce using its new disclosure avoidance system. This is an early opportunity for groups to see which statistical tables the bureau is proposing to produce and compare the proposed data products with those released from the 2010 Census.

OPPORTUNITIES FOR PROVIDING FEEDBACK

The bureau solicited feedback from the user community through a **Federal Register Notice** in 2018 and is expected to continue seeking input throughout 2019 and into 2020. In addition, the bureau provides regular updates on their plans for 2020 Census data products at **National Advisory Committee** (NAC) and **Census Scientific Advisory Committee** (CSAC) meetings. For each meeting, the bureau accepts public comments in person, by phone, or in writing (send comments to census.national.advisory.committee@census.gov and census.scientific.advisory.committee@census.gov for NAC and CSAC meetings respectively). The meetings are live-streamed and meeting materials may be found online (e.g., see **NAC 2019 fall meeting**).

There is also an opportunity for technical experts to weigh in. Since differential privacy allows for unprecedented transparency, the bureau is able to reveal how much noise is introduced into published statistics and even share the **algorithms and code** that are used. The bureau released a set of test data products in October 2019: this will be sample data of products based on the internal 2010 Census data and produced using the new disclosure avoidance system. These demonstration products are **available to the public**, and the bureau is inviting users to provide feedback at a **workshop** hosted by the National Academy of Sciences' Committee on National Statistics (CNSTAT) in December 2019. Users may also provide detailed feedback to the bureau at **dcmd.2010.demonstration.data.products@census.gov**.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of the following individuals for their input and feedback while producing this factsheet: Indivar Dutta-Gupta and Isabella Camacho-Craft of the Georgetown Center on Poverty and Inequality, danah boyd of Data & Society, Maria Filippelli of New America, Rosalind Gold of NALEO, Terri Ann Lowenthal, Terry Ao Minnis of Asian Americans Advancing Justice - AAJC, Kathy Pettit of the Urban Institute, Yvette Roubideaux of National Congress of American Indians, Arturo Vargas of NALEO, and Corrine Yu of The Leadership Conference on Civil and Human Rights. Please contact Jae June Lee (jl2435@georgetown.edu) and Cara Brumfield (cb1542@georgetown.edu) with any questions.