ASIAN AMERICANS ADVANCING JUSTICE AAJC
DATA& SOCIETY
MALDEF
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 & THEIR IMPLICATIONS

JAE JUNE LEE & CARA BRUMFIELD

KEY TAKEAWAYS

- The Census Bureau is changing how it protects the confidentiality of census responses. Confidentiality protections for the 2020 Census rely on a mathematical framework called "differential privacy."
- Differential privacy can provide robust and measurable confidentiality protections against evolving challenges presented by computing advances and the growing abundance and availability of data.
- Like past approaches, the new disclosure avoidance system involves a balance between confidentiality and the usefulness of data.
- Users of census data—including civil rights groups—have raised serious concerns about the impact the new disclosure avoidance system may have on the fitness-for-use of the data.

Every decade, the Census Bureau undertakes a vast and complex operation to count every person, once, only once, and in the right place. The agency is tasked with a dual mandate: it must produce useful statistics while ensuring that those statistics do not disclose, or allow others to discover, sensitive information about individual households. Starting with the 2020 Census, the bureau's Disclosure Avoidance System (DAS) will rely on a mathematical framework—called "differential privacy"—to manage confidentiality risks by introducing an adjustable amount of statistical imprecision or "noise" into statistics. Differential privacy offers robust and measurable bounds on confidentiality risks despite an unknowable amount of other data on the same households in the world now and in the future.

At the same time, the new DAS will affect the quality or fitness-for-use of census data. Alongside other sources of error during data collection and processing, the new DAS contributes to a lack of precision in resulting statistics. Users of census data have voiced serious concerns that the new DAS may unacceptably limit the utility of census data products—undermining an array of uses, including research, local policy decisions, and the enforcement of civil rights laws. Yet, stakeholders are also deeply aware of the serious confidentiality concerns among many of the communities they advocate for and represent. Stakeholders can and should weigh in and engage the bureau while the agency continues the development of the new DAS—likely through 2022—and determines the balance between confidentiality and data utility.

THE BUREAU'S CHANGING CONFIDENTIALITY PROTECTIONS

The bureau faces a fundamental challenge: publishing more statistics with greater detail and accuracy grows the probability that an external actor can reconstruct the underlying data and identify individual census households. Historically, the bureau employed a collection of methods to prevent disclosure of confidential information. This included "swapping" specific data elements, limiting the amount of published information, and replacing information about a respondent with a statistically estimated value using the "blank and impute" technique. 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 sources, present new evolving threats to confidentiality.

In response, the bureau has turned to differential privacy. While prior confidentiality protections were applied to specified elements of the data, the bureau will now introduce a programmable quantity of statistical imprecision, or "noise," to nearly 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 amount of noise to change the balance of utility and confidentiality. (The parameter that controls the balance of utility and confidentiality is referred to as the "privacy-loss budget" which is usually denoted using the Greek letter "ε" (epsilon). Put another way, published statistics are approximate—and not exact matches—to the underlying, confidential data, and this distance from underlying unpublished data can be measured and calibrated.

For the 2020 Census, the bureau is devising algorithms that satisfy the mathematical definition of differential privacy. However, the bureau's algorithms must meet additional constraints, through steps referred to as "post-processing," that are important for policy and technical goals beyond data confidentiality. These include requirements to produce counts as non-negative integers (in other words, a census block will not have negative or fractional counts of residents) or to hold certain counts as "invariants." Invariants such as the total population of a state will not be affected by the 2020 DAS and will be published as enumerated.

THE 2020 DAS & FITNESS-FOR-USE OF CENSUS DATA

As the bureau implements the 2020 DAS, the agency must consider how the myriad use-cases of these data will be impacted by the statistical noise introduced by disclosure avoidance. For example, these data are used in the enforcement of the Voting Rights Act. Overly imprecise race and ethnicity data could limit the ability of redistricting experts to create "majority-minority" districts that are designed to prevent the disenfranchisement of voters of color. In publishing census data, the bureau should balance the interests of census data uses, such as the enforcement of the Voting Rights Act, with the strong legal requirements for protecting the confidentiality of census respondents. While disclosure avoidance methods including differential privacy introduces statistical uncertainty into published statistics, there are other sources of error, including nonsampling errors, such as measurement and nonresponse errors (i.e. undercounts), that also contribute to a lack of precision in statistics.

Analyses have found that the imprecision introduced by post-processing is not uniformly experienced across subpopulations or geographic areas: typically, rural areas will see a greater relative distortions than urban areas, and numerically smaller subpopulations will also see greater relative swings in population counts than larger groups. (These systematic biases are a consequence of a variety of factors, including invariants and the greatly varying population sizes of census blocks.) This has led groups, such as the National Congress of American Indians, to raise concerns about substantial amounts of noise in statistics for small populations living in remote areas, potentially harming the quality of statistics about tribal nations.

STAKEHOLDER ENGAGEMENT WITH THE CENSUS BUREAU

The bureau's Data Stewardship Executive Policy (DSEP) committee—composed of senior career staff members—serves as the focal point for policy decisions regarding the design of the 2020 DAS. In November 2020, DSEP determined the final list of invariants. DSEP will also make decisions regarding the privacy loss budget to determine the balance between the utility and confidentiality of published statistics. Since successive data releases cumulatively increase confidentiality risks, DSEP must decide what the acceptable level of privacy-loss is across all decennial data products and determine how to best allocate this global privacy-loss budget across different data releases.

Since the 2020 DAS was announced, stakeholders have requested more robust communication and engagement with the bureau. To date, the bureau has engaged stakeholders through the Federal-State Cooperative for Population Estimates, the National Academies, and ongoing communications with a small group of civil rights organizations. The bureau is also holding 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 has also engaged the census advisory committees on the 2020 DAS, but unfilled vacancies have prevented one of the committees from relaying formal recommendations. The public can provide feedback to the bureau at the following email address: 2020DAS@census.gov.

The next DAS milestone will be the release of the Public Law 94-171 redistricting tables—the first set of data products processed through the 2020 DAS. While the statutory deadline for the delivery of the redistricting files is March 31, 2021 (as of February 2, 2021), news outlets have reported that the redistricting data will not be available until July 31, 2021 or later. Stakeholders are engaging the bureau on recommendations which include reducing the granularity of statistics and setting an adequately large privacy-loss budget for the redistricting file. DSEP will set the specific privacy loss-budget for the redistricting files in the summer of 2021 and decisions will be made during the course of 2021 and 2022 as the bureau publishes the full suite of expected data products and designs the differentially private methodologies for protecting these data products.