Why the Census Bureau Chose Differential Privacy

2020 Census Briefs

By the Population Reference Bureau and
the U.S. Census Bureau's 2020 Census Data Products and Dissemination Team

C2020BR-03

March 2023

This is the second in a series of briefs describing how disclosure avoidance methods are being applied to 2020 Census data products and implications of those methods for data users. More detailed information is available in the U.S. Census Bureau's handbook, "Disclosure Avoidance for the 2020 Census: An Introduction" and key points summarized in a brief, "Disclosure Avoidance and the 2020 Census Redistricting Data".[1]

WHY IS THE CENSUS BUREAU MODERNIZING PROTECTIONS FOR 2020 CENSUS DATA PRODUCTS?

The purpose of this brief is to explain how and why the Census Bureau applied a new disclosure avoidance system, based on differential privacy, to protect respondents' information in 2020 Census data products. This brief also highlights how the Census Bureau has engaged with data users while developing this new disclosure avoidance system. The decennial census is the premier source for information about America's changing population and households and a critical component of America's democracy. Decennial census data determine congressional apportionment, are used by states for redistricting, and inform the allocation of federal funding each year.

What Is Differential Privacy?

Differential privacy is a scientific framework for processing data to protect the identities and personal information of the people in the data. It works by adding *statistical noise*—small, random additions or subtractions—to every published statistic so that no one can reidentify a specific person or household with any certainty using any combination of the published data.

Differential privacy forms the foundation of the Disclosure Avoidance System used to adjust the data to protect 2020 Census respondent confidentiality.

The challenge for the Census Bureau is balancing the need to collect and report these data with the statutory obligation to protect respondent confidentiality.[2] The Census Bureau's work toward that balance is guided by a set of privacy principles that include necessity, openness, respectful treatment of respondents, and confidentiality.[3]

[1] U.S. Census Bureau, "Disclosure Avoidance for the 2020 Census: An Introduction," <www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html> and "Disclosure Avoidance and the 2020 Census Redistricting Data," <www.census.gov/library/publications/2023/decennial/c2020br-02.html>.

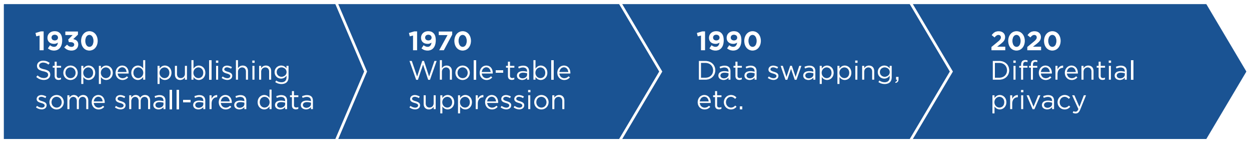
[2] U.S. Constitution, Article I, Section 2; Title 13 U.S. Code, Sections 8–9; Title 13 U.S. Code, Section 141.

[3] U.S. Census Bureau, "Our Privacy Principles," <www.census.gov/about/policies/privacy/data_stewardship/our_privacy_principles.html>.

United States® Census Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Figure 1.
Brief History of Census Disclosure Avoidance Protection



Source: U.S. Census Bureau.

WHICH 2020 CENSUS DATA ARE AFFECTED BY THE NEW DISCLOSURE AVOIDANCE SYSTEM?

It is against the law for the Census Bureau to disclose or publish any confidential information that identifies an individual or business, including names, addresses (including GPS coordinates), and telephone numbers. The decennial census does not ask for Social Security numbers, bank account information, religion, or information about political party affiliation.

The 2020 Census included just a small number of questions. Most of the responses to those questions—such as age—are used to publish statistics and are protected by disclosure avoidance methods. But not all question responses are included in published statistics. Names, dates of birth, and phone numbers are not published. This information is only used to help the Census Bureau meet their goal of “counting everyone once, only once, and in the right place.”

In the 2020 Census, people living in households were asked:

- Name (not published).
- Relationship to householder (included in published statistics).
- Sex (included in published statistics).
- Age (included in published statistics).
- Date of birth (not published).
- Hispanic origin (included in published statistics).
- Race (included in published statistics).
- Number of people living at the housing unit, which is tabulated.
- Information about additional people staying at address on April 1 (not directly published but used for quality control and nonresponse follow-up).
- Information about tenure, owner- or renter-occupied (included in published statistics).
- Telephone number (not published).

The Census Bureau conducts a separate group quarters operation to count people who live in places other than housing units such as correctional facilities, college or university student housing, or military quarters. Similar information is collected for people living in group quarters, excluding relationship to householder, and published group quarters statistics are also subject to disclosure avoidance methods.

DISCLOSURE AVOIDANCE IS NOT NEW

The 2020 Census redistricting data were the first 2020 Census data that were protected using differential privacy, but disclosure avoidance is not new. Figure 1 provides a summary overview of how census confidentiality protections have evolved from the 1930 Census to the 2020 Census.

Beginning with the 1930 Census, the Census Bureau stopped publishing certain tables for small geographic areas to protect respondents’ confidential data. Title 13, U.S. Code provides for the confidentiality of census data.[4] For the 1970 and 1980 Censuses, the Census Bureau did not publish certain tables based on the number of people or households in a given area.[5]

In 1990, the Census Bureau began using more sophisticated techniques, such as data swapping, to protect against disclosure. With data swapping, the Census Bureau injects *statistical noise* into the data by swapping the geographic identifiers on records for certain households with the identifiers from nearby households with similar characteristics. By design, the Census Bureau does not release information about its specific details for swapping. This is necessary to protect against disclosing individual information. As a result, the practice is not transparent to data users and

[4] U.S. Census Bureau, “Title 13 - Protection of Confidential Information,” <www.census.gov/about/policies/privacy/data_stewardship/title_13_-_protection_of_confidential_information.html>.

[5] U.S. Census Bureau, “Disclosure Avoidance Techniques Used for the 1970 Through 2010 Decennial Censuses of Population and Housing,” <www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20Techniques%20for%20the%201970-2010%20Censuses.pdf>.

they are unable to assess the impact of swapping on the accuracy of published data.

The Census Bureau continued to use data swapping in the 2000 and 2010 Censuses. It also used techniques such as top- and bottom-coding (grouping values above or below certain thresholds into broader categories), blank-and-impute techniques (replacing actual responses with statistically generated data), table and cell suppression (not publishing certain data points), and other methods to protect responses against disclosure.[6]

To modernize disclosure avoidance methods for the 2020 Census, the Census Bureau implemented a new framework based on differential privacy.

WHY IS THE CENSUS BUREAU USING NEW METHODS FOR CONFIDENTIALITY PROTECTION?

Advances in computing technology and rapid growth in the number of commercially available databases on people and households have increased concerns about data confidentiality. The number and complexity of decennial census publications have also increased over time, with the 2010 Census releasing about 150 billion U.S. population and housing statistics.[7]

Imagine a small neighborhood with only a few households where most people have the same characteristics. If one unique characteristic appears in that neighborhood's data, others may be able to easily guess the identity of the person with that characteristic.

The Census Bureau used older disclosure avoidance techniques to try to protect against these types of attacks in the past. Today, this type of disclosure risk can happen more easily—even for large geographies with lots of unique people—because of advances in technology. For example, researchers at the University of Washington showed how mathematical models could be used to reveal the identity of people who are transgender in previously published tables.[8, 9]

New types of disclosure risk call for new types of disclosure protection.

Reducing the Risk of Reidentification

Published tables from the Census Bureau are increasingly vulnerable to *database reconstruction and reidentification attacks*.[10] In database reconstruction, an outside party combines information in published tables and uses mathematical models to reconstruct the original census responses without names or addresses. In reidentification, an outside party links the reconstructed data to external databases (or uses personal knowledge about a person) on variables shared with the census responses and infers confidential information about individual census respondents.

The Census Bureau conducted an experiment to reconstruct a dataset of individuals using only published 2010 Census data tables. Note the data in these tables had undergone the 2010 data swapping disclosure avoidance application. The experiment resulted in reconstruction of a dataset of more than 300 million individuals. The Census Bureau then used that dataset to match the reconstructed records to four commercially available data sources, to attempt to identify the age, sex, race, and Hispanic origin of people in more than six million blocks in the 2010 Census.

The results of this database reconstruction simulation were concerning and the reconstructed data, if published, would have violated the approved confidentiality protection standards that were in place for all 2010 Census publications. The overall agreement rate—the percentage of reconstructed statistics that exactly match the geographic location, sex, age, race, and ethnicity of census respondents—was 91.8 percent (Figure 2).[11] Blocks with one to nine people had a lower agreement rate (74.0 percent) because households in the smallest blocks were the primary target of data swapping techniques in the 2010 Census, but slightly larger blocks with 10 to 49 people had an agreement rate of 93.0 percent.[12]

This simulation was based on a small subset of published data. The availability of vast databases of personal information at private companies multiplies the risk of reidentification, as this information could be disclosed through an online security breach

[6] U.S. Census Bureau, "Disclosure Avoidance Techniques Used for the 1970 Through 2010 Decennial Censuses of Population and Housing," <www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20Techniques%20for%20the%201970-2010%20Censuses.pdf>.

[7] John M. Abowd and Michael B. Hawes, "Confidentiality Protection in the 2020 US Census of Population and Housing," arXiv:2206.03524, <https://arxiv.org/abs/2206.03524>.

[8] Os Keyes and Abraham D. Flaxman, "How Census Data Put Trans Children at Risk," Scientific American, <www.scientificamerican.com/article/how-census-data-put-trans-children-at-risk/>.

[9] Travis Dick et al., "Confidence-Ranked Reconstruction of Census Microdata From Published Statistics," 2023, <https://doi.org/10.48550/arXiv.2211.03128>; Sallie Keller and John Abowd, "Database Reconstruction Does Compromise Confidentiality," 2023, <www.pnas.org/doi/10.1073/pnas.2300976120>.

[10] U.S. Census Bureau, "The Census Bureau's Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census,"<www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/simulated-reconstruction-abetted-re-identification-attack-on-the-2010-census.html>.

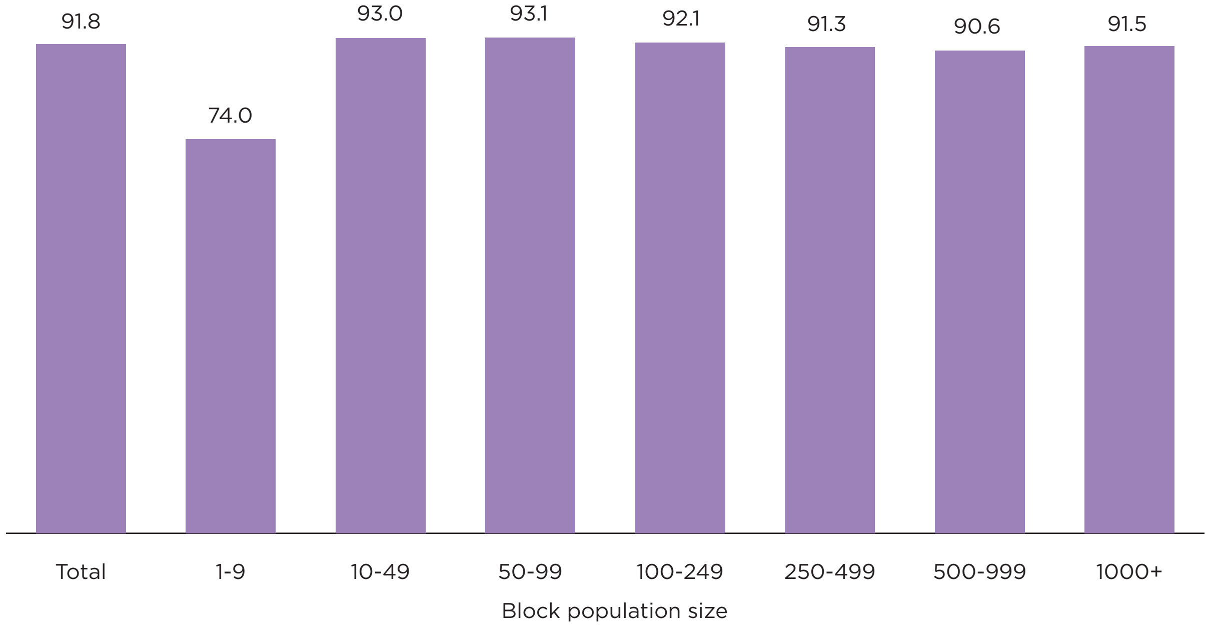
[11] John M. Abowd and Michael B. Hawes, "Confidentiality Protection in the 2020 US Census of Population and Housing," arXiv:2206.03524, <https://arxiv.org/abs/2206.03524>.

[12] Ibid.

Figure 2.

Agreement Rates of Reconstructed Data to Original, Confidential Data File by Block Population Size

(In percent)



Source: John M. Abowd and Michael B. Hawes, "Confidentiality Protection in the 2020 US Census of Population and Housing," arXiv:2206.03524, <https://arxiv.org/abs/2206.03524>.

or misused by those with privileged access to the information.

Risk of disclosure is particularly high for people with unique combinations of age and sex in their communities, called population uniques. Nationwide, roughly 150 million individuals—almost one-half of the population, have a unique combination of sex and single year of age at the block level. If an attacker found a match for a population unique in their reconstructed data, they could be very confident in the accuracy of their reidentification of that person's other characteristics such as race and ethnicity.[13]

When presented with the results of the simulated attack, the Census Bureau's Data Stewardship Executive Policy (DSEP) Committee realized that stronger disclosure avoidance methods would be needed for the 2020 Census.[14] Older disclosure avoidance methods, like data swapping, were not designed to defend against potential database reconstruction and reidentification attacks. If traditional disclosure avoidance methods were applied to the 2020 Census data, the amount of noise required to protect against these types of advanced attacks would almost certainly have made the census data unfit or too inaccurate for most uses. To provide adequate confidentiality protections and census data that are fit for use, the Disclosure Avoidance System had to be modernized.[15]

Advantages and Challenges of Differential Privacy

Differential privacy is a modern disclosure avoidance framework that provides mathematically provable confidentiality guarantees against a wide range of potential attacks. In other words, even if a new type of attack is developed, the math of differential privacy means that the data will be equally protected against future attacks. Differential privacy offers significant advantages over traditional approaches to disclosure avoidance:

[13] Elsayed A. H. Elamir and Chris J. Skinner, "Record-level Measures of Disclosure Risk for Survey Microdata, S^{3}RI Methodology Working Paper M04/02, Southampton Statistical Sciences Research Institute, Southampton, UK, 2004, <https://eprints.soton.ac.uk/8175/1/8175-01.pdf>.

[14] Data Stewardship Executive Policy Committee, "Final DSEP Meeting Record," U.S. Census Bureau, Washington, DC, 2018, <https://www2.census.gov/about/policies/foia/records/disclosure-avoidance/appendix-g-dsep-meeting-record_2018-02-15.pdf>.

[15] U.S. Census Bureau, "2020 Census Disclosure Avoidance System Development and Release Timeline," <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/das-development-timeline.pdf>.

- Differential privacy allows the Census Bureau to track and address potential disclosure risk as it accumulates across each successive data release. This ability to track the balance between confidentiality protection and data accuracy is what separates differential privacy frameworks from most traditional methods of disclosure avoidance.
- Unlike prior methods of table suppression or record swapping, differentially private data can be published, analyzed, and combined with other data without any increased risk of disclosure. Once the data have been processed using differential privacy protections, there is no more privacy loss regardless of how the data are used. This means the published data will be protected from the time they were released until the National Archives releases the confidential records on April 1, 2092.[16]
- Differential privacy allows the Census Bureau to be transparent about the disclosure avoidance being implemented and the documentation is available to the public, unlike prior data protection methods such as data swapping. The Census Bureau has made the programming code, settings, and summaries of the noise and bias available to the public.
- Compared to other disclosure avoidance methods currently available (such as data suppression and data swapping techniques), differential privacy provides the best combination of confidentiality protections, data accuracy, and availability (Table 1).

[16] U.S. Census Bureau, "The "72-Year Rule," <www.census.gov/history/www/genealogy/decennial_census_records/the_72_year_rule_1.html#>.

Table 1.
Characteristics of Different Disclosure Avoidance Methods

Disclosure Avoidance Method	Implications for Census Data		
	Confidentiality	Accuracy	Availability
Data suppression	Rule-based suppression systems, when applied to individual tables, often fail to account for complex interactions between related tables and may not effectively protect confidentiality.	Missing data can generate biases when the published data are analyzed.	Data users were dissatisfied with the amount of suppression required in the 1980 Census, which led to the adoption of noise infusion via record swapping to protect confidentiality in the 1990–2010 Censuses.
Data swapping	The relatively low swapping rate used in the 2010 Census does not protect respondent confidentiality, and even very high swapping rates would have limited ability to protect against reidentification attacks.	Results of swapping experiments showed this method can lead to significant distortions in population and race counts and in age structure.	Swapping does not limit the availability of data, just the accuracy and confidentiality of the data.
Differential privacy	Provides mathematically provable measures of protection.	The Census Bureau defined accuracy targets for the redistricting data by working with both internal and external data users, and then conducted extensive analyses to set the parameters necessary to meet those targets.	In principal, does not limit the amount of data that can be made available, but releasing more data increases the amount of noise necessary for disclosure protection, which could impact the usability of the tables. Therefore, the Census Bureau made the decision not to release some tables.

Source: U.S. Census Bureau, "Comparing Differential Privacy With Older Disclosure Avoidance Methods," <www.census.gov/library/fact-sheets/2021/comparing-differential-privacy-with-older-disclosure-avoidance-methods.html>; "Research into Alternatives to Differential Privacy," <www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/research-into-alternatives-to-differential-privacy.html>; John M. Abowd and Michael B. Hawes, "Confidentiality Protection in the 2020 U.S. Census of Population and Housing," arXiv:2206.03524, <https://arxiv.org/abs/2206.03524>; and John Abowd et al., "The 2020 Census Disclosure Avoidance System TopDown Algorithm," Harvard Data Science Review (Special Issue 2), <https://doi.org/10.1162/99608f92.529e3cb9>.

However, as with any disclosure avoidance method, there are limitations to and challenges with the differential privacy approach. For example:

- Data for very small demographic groups and geographic areas, such as census blocks, may be too noisy for a particular use and should be aggregated into larger geographic areas before use.
- Noise infusion results in some implausible results—such as a block with more occupied housing units than people to occupy those units.
- The complexity of the methods makes it difficult to communicate how disclosure avoidance works and what it means for their data applications.

More information on challenges can be found in "Disclosure Avoidance and the 2020 Census Redistricting Data".[17]

HOW IS THE CENSUS BUREAU ENGAGING WITH DATA USERS?

External data users have provided essential feedback to the Census Bureau during the development of new disclosure avoidance methods for the 2020 Census.[18]

The Census Bureau's Data Stewardship Executive Policy Committee relies on input from a variety of sources when making decisions about the adoption, implementation, and settings of disclosure avoidance methods. These include internal subject matter experts, the Census Bureau's advisory committees, the Committee on National Statistics of the National Academy of Sciences, academic experts and researchers, privacy advocates, professional associations, federal and state partners, and American Indian and Alaska Native tribal leaders.

Between July 2018 and August 2021, the Census Bureau received over 1,200 public comments on 2020 Census data products. These comments informed the design of the data products and implementation of the disclosure avoidance system that was applied to the 2020 Census redistricting data tables.[19] Detailed information about these methods is provided in another brief, "Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works."[20]

Data user feedback has also been incorporated in a series of demonstration products to help data users evaluate whether the noise-infused data are fit for use.[21] Advanced data users may download demonstration data that were generated by applying the 2020 Census disclosure avoidance methods to the 2010 Census data. This allows side-by-side analysis of the 2010 published data against 2010 data with differential privacy applied. Keep in mind that this comparison is imperfect because the 2010 published data has the swapping method of disclosure avoidance applied to it.

Data users can analyze the Detailed Summary Metrics, which provide measures of error introduced by differential privacy for a range of topics, such as race and ethnicity, and geographies such as counties and census tracts.

Refer to the Census Bureau handbook on "Disclosure Avoidance for the 2020 Census: An Introduction" to learn how disclosure avoidance methods were applied to the 2020 Census redistricting data and what the implications are for data users.[22] Future briefs will address methods used for other 2020 Census data products.[23]

[17] U.S. Census Bureau, "Disclosure Avoidance and the 2020 Census Redistricting Data," <www.census.gov/library/publications/2023/decennial/c2020br-02.html>.

[18] Feedback received to date can be accessed by visiting Round 1 feedback, 2010 Demonstration Data Demographic and Housing Characteristics File (DHC) v. 2022-03-16 (6/23/2022), <https://www2.census.gov/programs-surveys/decennial/2020/program-management/round_1_feedback.pdf>, and Round 2 Feedback, 2010 Demographic and Housing Characteristics File (DHC) v. 2022-08-25 (11/8/2022), <https://www2.census.gov/programs-surveys/decennial/2020/program-management/round_2_feedback.pdf>.

[19] John M. Abowd and Michael B. Hawes, "Confidentiality Protection in the 2020 US Census of Population and Housing," arXiv:2206.03524, <https://arxiv.org/abs/2206.03524>.

[20] U.S. Census Bureau, "Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works," <www.census.gov/library/publications/2023/decennial/c2020br-04.html>.

[21] U.S. Census Bureau, "Developing the DAS: Demonstration Data and Progress Metrics, Detailed Summary Metrics for Production Settings," June 8, 2021, <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html>.

[22] U.S. Census Bureau, "Disclosure Avoidance for the 2020 Census: An Introduction," <www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html>.

[23] Ibid.

WHERE CAN I LEARN MORE?

- Disclosure Avoidance and the 2020 Census Redistricting Data <www.census.gov/library/publications/2023/decennial/c2020br-02.html>
- Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works <www.census.gov/library/publications/2023/decennial/c2020br-04.html>
- Disclosure Avoidance for the 2020 Census: An Introduction <www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html>.
- Disclosure Avoidance: Latest Frequently Asked Questions <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates/2020-das-faqs.html>
- 2020 Decennial Census: Processing the Count: Disclosure Avoidance Modernization <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html>
- Disclosure Avoidance Webinar Series <www.census.gov/data/academy/webinars/series/disclosure-avoidance.2021.List_882320526.html#list-tab-List_882320526>

You can also subscribe to the Census Bureau's "2020 Census Data Products Newsletter" for timely updates and contact us at <2020DAS@census.gov> if you have questions.[24]

[24] U.S. Census Bureau, "Decennial Census: Data Products and Operational Updates," <https://public.govdelivery.com/accounts/USCENSUS/signup/15409>.