Consistency of Data Products and Formal Privacy Methods for the 2020 Census

Contact: Gordon Long — glong@mitre.org

JSR-21-02

January 2022

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 983-6997

Contents

Abstract

The Census Bureau must balance the need to provide high quality data at fine granularity with its obligation to avoid disclosing sensitive information. In response to demonstrated reconstruction and reidentification attacks enabled by modern computing, the Census Bureau concluded that its traditional disclosure avoidance mechanisms were insufficient for the 2020 decennial census and adopted a modernized approach designed to achieve formal privacy guarantees. The methods used involve adding sampled random noise to results, which can introduce inconsistencies in the resulting data. JASON was asked to study the impact of these privacy mechanisms on consistency for two data products produced from the 2020 census data: the PL94-171 data produced for redistricting and the Demographic and Housing Characteristics file. This report provides findings and recommendations on the mechanisms adopted by the Census Bureau to achieve consistency, and the challenges of balancing utility and privacy for future Census Bureau data products.

1 EXECUTIVE SUMMARY

Every ten years since 1790, as prescribed by Article 1 of the United States Constitution, the United States Census Bureau conducts an enumeration of each person in the United States. The 2020 census marks the twenty-fourth decennial census, preceded by more than a decade of planning including extensive experimentation and tests. The authority to collect and analyze the information gathered by the Census Bureau originates in Title 13 of the U.S. Code enacted into law in 1954. The decennial census is responsible for more than just reapportionment and redistricting. The data products generated from the census are also used for disbursement of federal funds and are a critical source of data for policy makers and researchers. Two data products of particular focus for this study are the redistricting data known as the PL94-171 data (after the public law that describes it) which includes population counts by race and voting age down to the level of Census blocks, and the Demographic and Housing Characteristics (DHC) File, which provides statistics on household composition and attributes.

Title 13 prohibits the Census Bureau from making "any publication whereby the data furnished by any particular establishment or individual under this title can be identified" (United States Code, 1990). This presents the Census Bureau with a seemingly impossible task: to provide billions of statistics that accurately and comprehensively capture the American population, while not allowing any sensitive information about an identifiable individual to be disclosed from that data.

Following the potential risk of reidentification revealed by the Census Bureau's reconstruction and reidentification experiments on 2010 Census data, the Census Bureau concluded that previous disclosure avoidance methods resulted in data releases that could violate Title 13 requirements. The Census Bureau then embarked on a bold and ambitious plan to ensure that data releases from the 2020 Census would satisfy formal privacy guarantees that provide mathematical bounds

on disclosure risk. These guarantees are based on a mathematical definition of privacy known as differential privacy. Differential privacy has emerged as the gold standard for measuring privacy in this era where potential adversaries have access to powerful computers and extensive auxiliary information.

This effort has resulted in the development of novel algorithms for achieving formal privacy guarantees for Census data releases. Unlike the disclosure avoidance methods used for previous decennial censuses, the methods adopted for the 2020 census are transparent and the Census Bureau has openly discussed their design and implementation. Confidentiality depends on principled application of noise sampled from known distributions—no secrecy about the methods is needed, other than the actual random samples used. The Census Bureau conducted experiments to understand the impact of the mechanisms for achieving a formal privacy guarantee on data accuracy. They carried out a public process through a series of demonstration data releases based on the 2010 decennial census data to gather feedback from stakeholders on the impact of privacy mechanisms on data quality. This process led to changes in the privacy mechanisms, strategy for allocating privacy loss budget across geographic levels and queries, and an overall increase in the privacy loss budget.

Much has been learned from these efforts, including clarity on the significant challenges in meeting the competing goals of providing high quality data of the scale and comprehensiveness expected from the census while ensuring a meaningful formal privacy guarantee. Versions of these algorithms have been applied to the 2020 decennial census PL94-171 redistricting data that was released in August 2021 and are planned to be used for subsequent data products including the DHC data.

The modernized disclosure avoidance methods adopted by the Census Bureau raise concerns about consistency since formal privacy is satisfied by adding sampled random noise to statistics. Desirable consistency properties that may be violated by adding privacy noise include both formal consistency (being free

from internal contradictions) and semantic plausibility (e.g., producing population counts that are not negative).

The Census Bureau asked JASON to respond to the following questions:

- Is consistency between the PL94-171 data and the DHC data achievable?
- Is consistency between the PL94-171 data and the DHC data necessary?
- What recommendations does JASON have on how to communicate if we cannot achieve consistency?
- What recommendations does JASON have for the 2030 Census and other data products with universe household frames (e.g., administrative record censuses)?

The Census Bureau's implemented production system for the creation of both the PL94-171 and DHC data products uses common microdata at the level of an individual person or household to build the tables, so the answer to the first question is "yes". The decision to produce these data products from synthetic microdata both requires and ensures formal consistency, but the approach taken to achieve this may have impact on accuracy and bias of the resulting data. JASON's study of these issues resulted in the following findings and recommendations.

1.1 General Findings

Findings:

F1 The Census Bureau has taken advantage of research advances, and their own algorithmic innovations, to control utility losses associated with achieving a formal privacy loss guarantee.

F2 The theoretical basis for the privacy methods that are used is sound. However, the effect on disclosure risk of these methods, as implemented with the selected parameters, is not well quantified.

F3 The Census Bureau sought to satisfy utility needs while minimizing formal privacy loss, but concrete disclosure risks are not sufficiently quantified to factor into decisions about disclosure avoidance options.

F4 The Census Bureau is producing the PL94-171 and DHC from generated microdata as a result of internal operational requirements. The production from microdata approach imposes otherwise unnecessary constraints that impact data quality.

1.2 Privacy and Utility

Findings:

F5 Census data users express concerns about data inconsistencies, but the problems associated with inconsistent data could be resolved by the Census Bureau providing guidelines for users to follow when working with inconsistent data.

F6 Block level data are not needed for the main use cases of the DHC data.

F7 Block level data (and other highly detailed data, such as age-by-year) pose the greatest reconstruction-reidentification risks.

F8 Tribal lands have different requirements including needs for highly detailed data for areas with low population.

F9 The Census Bureau has optimized its geographical hierarchy to improve the accuracy of statistics for politically important geographic areas that do not correspond to traditional on-spine entities.

F10 The detailed queries and consistency constraints enforced in producing the post-processed data are required to produce suitable microdata, as needed for the Census production system.

- The Census Bureau uses a set of detailed queries with 2,016 cells for each geographic unit (Household or Group Quarters Type [8 values] × Voting Age [2] × Hispanic/Latino origin [2] × Race [63]) down to the block level (5,892,698 populated blocks) to produce the PL94-171 data product. The statistics corresponding to these queries are not directly included in the PL data release or any future planned data releases.
- The detailed queries consume a large amount of the formal privacy budget allocated to the PL release.
- The post-processing performed to ensure non-negativity introduces bias in the results.

F11 Without access to all the noisy measurements used to produce a published value, it is difficult for users to understand how published values relate to the enumerated values.

F12 The threats and risks to both society and Census Bureau reputation from inferential disclosure attacks on Census data have not been meaningfully quantified.

F13 It is unclear if the privacy mechanisms adopted are sufficient to mitigate the vulnerabilities.

Recommendations:

R1 The Census Bureau should not prioritize consistency, either within or across data products.

R2 The Census Bureau should minimize the characteristics that are released at block level and avoid releasing DHC data at the block level.

R3 The Census Bureau should release the noisy measurements corresponding to published tables. The Census Bureau should clearly justify any use of privacy loss budget on noisy measurements that do not correspond to published statistics.

R4 The Census Bureau should release data products using the optimized block groups. These are the units with the most accurate statistics, and users should be encouraged to use the optimized block groups instead of the traditional tabulation block groups except when historical continuity is required.

R5 The Census Bureau should (i) establish standards for acceptable inferential disclosure risks, (ii) conduct experiments to understand inferential disclosure risks associated with data releases, and (iii) publish results from these experiments. JASON recommends that the Census Bureau should:

(a) Conduct experiments that make the disclosure risks concrete. For example, quantifying the ability to infer race for individuals who are of non-modal race for their block or to find cohabiting couples with children.

(b) Study the impact of privacy parameters on disclosure risk.

(c) Conduct experiments to estimate the impact of suppressing selected data such as not releasing block-level data in the DHC.

(d) Conduct experiments to simulate worst-case attacks including creative attacks that do not just perform a reconstruction followed by a reidentification, and experiments involving simulated data with high-risk properties.

R6 The Census Bureau should conduct and publish results from experiments to better understand the impact of post-processing on the accuracy and biases of computed estimates.

R7 The Census Bureau should convene meetings with tribal representatives and

consider providing additional data to sovereign tribal governments in ways that satisfy their needs and recognize their distinct status.

1.3 Communications

Findings:

F14 The Census Bureau has put commendable effort into communicating about differential privacy and has engaged transparently with their stakeholders throughout the process of developing disclosure avoidance mechanisms for the 2020 census but has struggled to convince stakeholders that the selected methods appropriately balance utility and privacy.

F15 The most important communications the Census Bureau does are through its public data products.

F16 Differential privacy mechanisms introduce statistical features into the data that may be unfamiliar to data users.

F17 The Census Bureau plans to take measures to avoid releasing negative population counts in upcoming data products, partly because of fears that negative values would be confusing and problematic to users. Including negative values poses communications challenges, but also provides an opportunity to clearly communicate the impact of privacy noise. Requiring non-negativity introduces bias and conceals the presence of privacy noise and complicates the methods the Census must communicate to users.

F18 Many users of Census data use widely available software tools for data analysis. Software vendors could adapt their tools to process annotated, noisy measurements and to produce more useful results from the provided data, including estimates of uncertainty.

Recommendation:

R8 The Census Bureau should not reduce the information value of their data products solely because of fears that some stakeholders will be confused by or misuse the released data.

(a) When possible, without unduly increasing disclosure risk, all noisy measurements that are used to produce a published statistic should be released, and the process used to produce published data should be transparent and reproducible.

(b) Data releases should include explicit information on the privacy noise distribution used for each cell and any post-processing.

(c) Data releases should include estimates of all sources of uncertainty.

R9 Concurrently with releasing the noisy measurements, the Census Bureau should provide post-processed statistics, along with reproducible programs that generate the official post-processed statistics from the noisy measurements.

R10 The Census Bureau should engage with statisticians and developers of statistical software commonly used on Census data (e.g., R Consortium, Microsoft Excel, SAS, SPSS, and Stata) to develop methods for working with annotated, noisy measurements and incorporating these into software tools.

1.4 Looking to the Future

Findings:

F19 The current interpretation of Title 13's non-identification requirements is incompatible with modern technical understanding of privacy.

F20 Much has been learned about the costs and complexity of achieving formal privacy for data releases and satisfying microdata and consistency requirements, but not enough is known about whether the privacy mechanisms as implemented are sufficient to mitigate the disclosure risks that motivated adoption of formal privacy.

Recommendations:

R11 The Census Bureau should seek clarification of, or modification to, the Title 13 confidentiality requirements and plan the 2030 Census around an operational and achievable interpretation of Title 13.

R12 The Census Bureau should take an approach to the 2030 census that builds upon what has been learned from 2020, starting with developing and articulating concrete disclosure avoidance requirements for the 2030 Census data releases and designing disclosure avoidance mechanisms and data products to provide maximum utility while satisfying those requirements.

This Page Intentionally Left Blank

2 INTRODUCTION

Every ten years since 1790, as mandated by the United States Constitution, the Census Bureau conducts a complete enumeration of all persons living in the United States. The 2020 census marks the twenty-fourth decennial census, and involved more than a decade of planning, including extensive experimentation and tests, leading to a set of public data products. The decennial census is responsible for providing population data for reapportionment and the production of a national public good information platform through the dissemination of its data products.

The Census Bureau's authority to collect the information and release data products derived from it originates in Title 13 of the U.S. Code, enacted into law in 1954. Federal law specifies that the census be accurate while simultaneously protecting the privacy of all individuals whose data is included. In particular, Section 9 of Title 13 mandates that information collected by the Census Bureau must be treated as confidential, and that the Census may not "make any publication whereby the data furnished by any particular establishment or individual under this title can be identified".

To satisfy these confidentiality requirements, the Census Bureau has used a variety of disclosure avoidance methods in past censuses (McKenna, 2018), including table and cell suppression, controlled rounding, and swapping records between blocks. In light of advances in algorithms and computing power, however, these methods were found to be insufficient. Consequently, new disclosure avoidance methods were developed for the 2020 census. These methods are designed to achieve a formal notion of privacy that provides mathematically established limits on any inference attacks on the released data, regardless of the auxiliary information and computing power available to an adversary.

This report analyzes the impact of the disclosure avoidance methods that have been adopted for the PL94-171 data product and that are planned for use in the upcoming Demographic and Housing Characteristics file (DHC) release. The

report considers the impact of the disclosure avoidance methods used by the Census Bureau on privacy and utility, with a particular focus on consistency. We start by providing background on the data products the Census Bureau produces from the decennial census data, introducing the tension between disclosure avoidance and consistency, and overviewing the process JASON followed to conduct the study.

2.1 Decennial Collection and Data Products

The goal of the census enumeration is to create a complete and accurate list of every resident in the United States. The Census Bureau relies on addresses to identify the locations of housing units, including group quarters, and then enumerates the population at those locations. This is done by creating a Master Address File (MAF) prior to the enumeration and confirming the addresses in the MAF are complete and correct by canvassing the country through administrative records (government and private sector) and in-person where needed. The MAF was first created in 1970 and has been augmented and improved every decade since.

No names are attached to the MAF. Rather, the MAF is used as the basis for canvassing the country a second time to count the people associated with each housing unit represented on the MAF. For the 2020 census, this second canvassing used a variety of modalities to reach each housing unit and collect the data including self-response via mail or internet, personal contact by an enumerator, and administrative records.

The respondent data are collected as a list of records indicating the responses for each resident within each housing unit. The Census Bureau refers to the records associated with individuals as *microdata*. Quality controls are performed to de-duplicate the records and validate addresses to ensure that every person is counted only once and at the right location. It is at this stage that housing unit status is imputed if necessary. This forms the Census Unedited File (CUF). The state counts for apportionment are computed from the CUF.

The Census Bureau then produces the Census Edited File or the CEF. This is done by resolving missing and inaccurate demographic data in the CUF. To resolve conflicting data or fill in missing information multiple methods are used, including comparisons to previous census data and to administrative records or statistical imputations. This results in the sensitive CEF data. Before any data derived from the CEF can be released, confidentiality protections are applied via the Disclosure Avoidance System (DAS), which is the main focus of this study. The DAS produces privacy-protected synthetic microdata that can then be processed through their production system to produce the public use data products.

The Census Bureau produces many data products from the data collected for the decennial census. With the exception of the apportionment counts, these products are derived from the CEF. They include the PL94-171 redistricting data and the DHC file which are described in detail below. Other products include the Demographic Profile file, the Detailed DHC product which includes statistics for detailed race and ethnicity groups and complex person/household tables not included in the DHC, the Congressional District Demographic and Housing Characteristics File, the Public Use Microdata Sample (PUMS), Census Briefs, Census Population and Housing Tables (CPH-T), Special Reports, and Special Tabulations. The scope of this JASON study is limited to the PL and DHC data products, and we do not consider the other data products.

2.1.1 PL94-171 redistricting data summary file

U.S.C. Public Law (PL) 94-171, enacted in 1975, amends Section 141 of U.S.C. Title 13 regarding the obligations of the Census Bureau for supporting redistricting (United States Code, 1990, 1974). It directs the Secretary of Commerce to work with the states, starting four years before the upcoming decennial census year, to define sub-state geographies and tabulations to support legislative redistricting. States participate voluntarily and the Secretary is given the latitude to determine the form and content of these data including the use of sampling procedures and

Table 2-1: Tables in the PL94-171 Redistricting Data Summary File (the number of cells in each table attributes captures the size of the table, but is less than the number of cells in the tables, because the published tables include additional cells with redundant information) (United States Census Bureau, 2020a).

Table	Contents	Cells	Explanation
P1	Race	71	63 races, and redundant totals
P2	Ethnicity by Race	73	63 races × 2 ethnicities; published table is counts for "Hispanic or Latino, and Not Hispanic or Latino by Race", complement can be derived from P1
P3	Race for the population 18 years and over	71	63 races, and redundant totals
P4	Ethnicity by Race for the population 18 years and over	73	as done for P2
P5	Group quarters population by major group quarters type	10	7 group quarters types, and redundant totals
H1	Occupancy status	3	2 (either occupied or vacant) statuses, and redundant total

special surveys. In recent history, the Census Bureau has provided data for a variety of geographic areas within a state, including tabulations that used block-level decennial enumerations. The data released to satisfy this law are known as the PL94-171 Redistricting Data Summary File, which we will refer to as the "PL" data. The 2020 PL data were released on August 12, 2021.

For the 2020 census, the PL data comprises the six tables listed in Table 2-1. Each cell in the table is a positive integer that gives the count of the number of people in the geographic unit who satisfy the characteristic. The categories for race and ethnicity are prescribed by the Office of Management and Budget (1997). There are 63 race categories comprising all combinations of the six designated race categories (none is not an option, hence $2^6 - 1$ race categories): White, Black or African American, American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander, and Some Other Race. There are two ethnic-

ities: Hispanic or Latino, and Not Hispanic or Latino. The PL tables also include population counts for Group Quarters, which comprises the institutionalized population (correctional facilities for adults, juvenile facilities, nursing facilities, and other institutional facilities) and noninstitutionalized population (university student housing, military quarters, other noninstitutional facilities). These tables are produced from the *Persons* file, which contains a microdata record for each individual enumerated by the census. The PL data includes one additional table, H1, which captures the number of occupied and vacant housing units. This table is produced from the *Housing Units* file which contains data on number and status of housing units.

Tabulations are provided at various levels of geography. The standard hierarchy of Census geographic entities is given in Figure 2-1. The line through the center of the figure is called the census *spine*. The P1–P5 and H1 tables are provided for every geographic unit on the spine, down to the block level. The *spine* geographic units are summarized in Table 2-2.

The PL tables include some semantically redundant data—for example, the total population of the geographic unit is the sum of the population counts for each race category. It also contains some redundant information across the geographic levels—the total population of a county is the sum of the populations of each of the census tracts it contains. We discuss these, and other consistency issues, in Section 2.2.

2.1.2 Demographic and housing characteristics file (DHC)

The Demographic and Housing Characteristics (DHC) File consists of more detailed demographic information than the PL, including tables for age-by-year and householder relationships. The DHC will include detailed geographic tabulations with some starting at the Block level and others starting at the Tract, County, and State level. An illustrative sample of the tables that are planned for inclusion in the

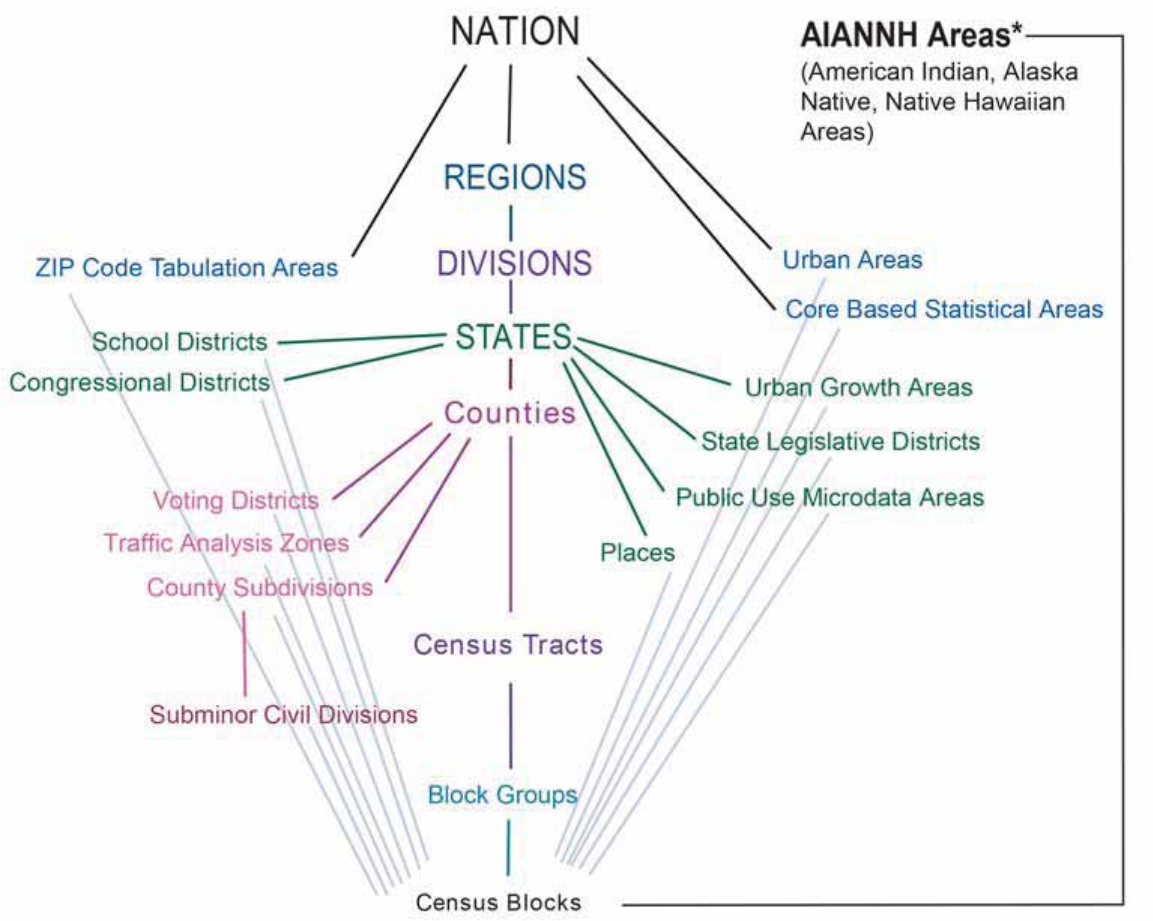

Figure 2-1: Standard hierarchy of Census geographic entities (United States Census Bureau, 2020b).

DHC are summarized in Table 2-3. The DHC is tentatively planned to be released in 2022.

2.2 Disclosure Avoidance and Consistency

The Census Bureau has long been concerned with confidentiality. In addition to their legal requirements to protect the privacy of individual, the Census Bureau needs to engender the trust of the public because they depend heavily on public cooperation. Due to concerns that the traditional disclosure avoidance mechanisms used in the 2010 census and earlier censuses were not sufficient to provide

Table 2-2: Geographic Units in 2020 Census, as used in the Disclosure Avoidance System (Abowd et al., 2021; United States Census Bureau, 2018)

Level	Number	Notes
Nation	1	
State	88	Includes DC, Puerto Rico, and state-level American Indian, Alaska Native, Native Hawaiian Areas
County	3,496	Includes parishes (Louisiana), independent cities (Maryland, Missouri, Nevada, and Virginia), municipios (Puerto Rico), and some other districts
Tract	84,589	
Block Group	239,780	Tabulation block groups; Section 5.2 explains how block groups were optimized for the 2020 census
Block	5,892,698	Blocks that do not contain any housing unit or occupied group quarters are not included

privacy in light of the vulnerabilities enabled by modern computing, the Census Bureau adopted a new method for disclosure avoidance for the 2020 census data products. This method is designed to achieve a formal notion of privacy known as *differential privacy*, and involved adding randomly sampled noise to published statistics to provide formal privacy guarantees that limit what any adversary could infer from the released data. Differential privacy mechanisms provide an inherent and quantifiable tradeoff between the amount of privacy provided (measured by the *privacy loss budget*) and the loss of data utility due to the scale of the noise required. We provide background on disclosure avoidance methods and differential privacy in Section 4. The Census Bureau was one of the first organizations to release an application using differential privacy mechanisms (Machanavajjhala et al., 2008), and its plans to use differential privacy for the 2020 census have been celebrated by the academic privacy community, widely reported in the mainstream press, but severely criticized by certain stakeholders (boyd, d, 2021).

The main issue that is the focus of this study is that the noise added may produce statistics violating expected consistency properties. We consider three distinct types of consistency:

Table 2-3: Selection of Tables planned for the DHC (United States Census Bureau, 2020a). (Tables selected from DHC Crosswalk list to give an incomplete, but representative, sense of the level of detail and comprehensiveness of the planned DHC release.)

Table	Contents	Granularity	Cells	Explanation
P7	Sex by Age	Block	49	23 age categories × 2 genders (and totals)
P7[A–I]	Sex by Age for each Race	Block	49	same as P7, separate table for each of the 6 race categories alone, and for two or more races, Hispanic, and white alone not Hispanic
PCT1	Sex by Age (by year)	Tract	209	103 age categories (by year up to 100) × 2 genders
P8	Median Age by Sex	Block	3	decimal number × 2 genders and both
P9	Sex by Age for Population Under 20 Years	Block	43	20 single year ages × 2 genders
P12	Households by type and presence of own children under 18	Block	19	4 household types (married couple, cohabiting, single male, single female); with and without children and relatives.
H10	Tenure by Race of Householder	Block	17	Owner or Renter × 7 race categories (6 races alone, two or more)
H14	Tenure by Household Type by Age of Householder	Block	69	Categories broken down by type at 3 age levels (e.g., renter, non-family, single female, 35 to 64 years)
PCO13	Group Quarters Population by Sex by Age	County	39	18 age categories × 2 genders, totals
PCO20	Group Quarters Population in Student Housing by Sex by Age	County	13	5 age categories × 2 genders, totals

- *Formal consistency* — data should be free from internal contradiction. This means any way of computing semantically equivalent values from the data should produce the same result. So, for example, the value of a statistic that gives the total population of a county is equal to the sum of the populations of all the census tracts that comprise that county.
- Semantic consistency (*plausibility*) — data do not violate properties that are expected to hold because of the meaning of the data. For example, population counts are never negative and parents cannot be younger than their biological children.
- Consistent with reality (*accuracy*) — data are as close as possible to the true value.

The formal consistency notion can be precisely defined, as long as it is clear what values are semantically equivalent. For the example above, equivalence depends on knowing that every country is partitioned into a set of census tracts and every enumerated individual in a county must be in exactly one of those census tracts. The Census Bureau and users of its data products have traditionally expected all data products released from the same underlying microdata to have strong formal consistency.

Plausibility depends on common understanding of the semantics of particular statistics. In some cases, the meaning of a statistic may be precisely defined in a way that makes certain values implausible. For example, occupied housing units are defined such that their population must be at least one. In most cases, what makes statistics implausible can be somewhat subjective. For example, most would consider it implausible for a block to have many children under 12 but no adults, but there would be less agreement on the implausibility if the population considered is under 18.

Consistency with reality is difficult to measure since, for most of the statistics the Census Bureau publishes, the ground truth value is unknown. When analyzing

privacy-protected statistics, the term *accuracy* is often used to evaluate how close the privacy-noised data are to the underlying enumerated values. However, privacy noise is only one of the several possible sources of inaccuracy in the published official statistics. We discuss efforts the Census Bureau has undertaken to estimate overall accuracy in Section 6.1.

2.3 JASON Study Process

The Census Bureau asked JASON to respond to these four questions:

- *Is consistency between the PL94-171 data and the DHC data achievable?*
- *Is consistency between the PL94-171 data and the DHC data necessary?*
- *What recommendations does JASON have on how to communicate if we cannot achieve consistency?*
- *What recommendations does JASON have for the 2030 Census and other data products with universe household frames (e.g., administrative record censuses)?*

We interpreted *consistency* in these questions as primarily meaning the notion of formal consistency as defined in the previous subsection, but also considered the plausibility and accuracy forms of consistency. In particular, the methods the Census Bureau adopted to support plausibility by avoiding negative counts in released statistics interacts with the other desired properties in important ways, and understanding overall accuracy is important for evaluating the impact of privacy noise.

Figure 2-2 summarizes the scope of the study. Section 3 provides JASON's analysis of the uses of decennial census data. Section 4 provides background on formal privacy and Section 5 describes the modernized disclosure avoidance methods used for the 2020 census. We provide JASON's evaluation of these

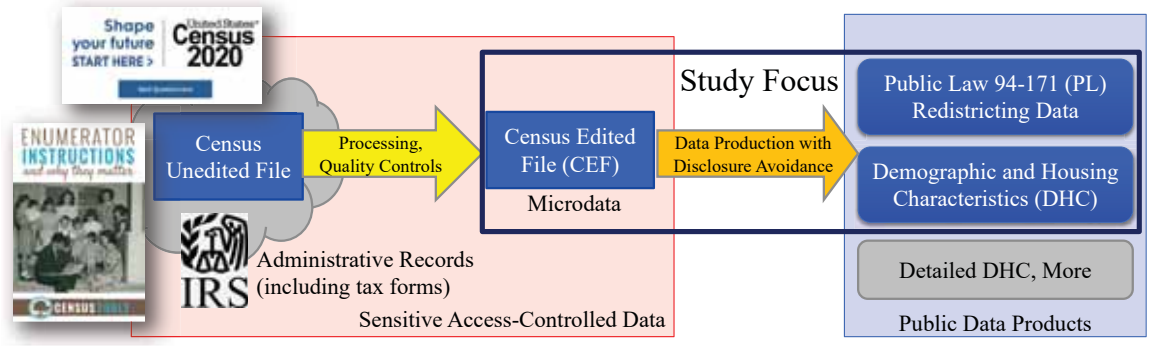

Figure 2-2: Scope of this study. The focus of the study is on the process used to produce the public PL94-171 and DHC data products from the Census Edited File, and the impact of privacy mechanisms adopted for the 2020 census on those data products.

methods and their impact on consistency in Section 6. In Section 7, we provide and support recommendations on communicating with Census Bureau stakeholders. Section 8 considers future Census Bureau data releases, including planning the 2030 decennial census. In Section 9, we summarize JASON's findings and recommendations.

JASON was introduced to the relevant issues during in-briefings arranged by the Census Bureau that were held in La Jolla, California on 14–15 June 2021. The presentations given during these briefings are listed in Table 2-4. Due to Covid-19 travel and meeting restrictions, these briefings were done virtually. The in-briefings also included discussions with John Abowd (Chief Scientist of the Census Bureau), Michael Hawes (Senior Advisor for Data Access and Privacy, Census Bureau), and Ron Jarmin (Acting Director of the United States Census Bureau) who attended the in-briefings in person.

Following the in-briefings, the Census Bureau arranged a series of virtual briefings with technical members of the DAS Team including Michael Hawes, Philip Leclerc, Ryan Cumings, Scott Holan, Ryan Janicki, Theresa Nguyen, Pavel Zhuravlev, Ashwin Machanavajjhala (Duke University and Tumult Labs), Daniel Kifer (Pennsylvania State University), and subsequent follow-up discussions by email and video-conference with Philip Leclerc and Michael Hawes.

Table 2-4: Study In-Briefings (14–15 June, 2021)

Briefer	Topic
Jason Devine (Census Bureau)	Census Data Products
Robert Ashmead (Census Bureau)	Top Down Algorithm and Privacy Semantics
James Whitehorne and Jason Devine (Census Bureau)	Link Between Redistricting Data and DHC
Christine Hartley (Census Bureau)	Population Estimates Program
Matt Spence (Census Bureau)	Summary Metrics and DHC Use Cases
Yvette Roubideaux (National Congress of American Indians)	American Indian/Alaska Native Use Cases for 2020 Census Data
Jeff Hardcastle (Nevada State Demographer)	Winners and Losers with the Privacy Loss Budget
Jan Vink (Cornell University)	DHC Accuracy and Use
Qian Cai (University of Virginia)	Data Consistency between the 2020 Census PL and DHC Data
Mike Mohrman (Washington State)	Washington's Small Area Estimates
Chris Dick (DA Advisors)	DHC Uses: Equity, Health, and Government

Table 2-5: Additional Experts Consulted by JASON

Briefer	Topic
Cynthia Dwork (Harvard University)	Differential Privacy
Joseph Salvo (Former Chief Demographer at NYC Department of City Planning)	Use of Census Data
Jonathan Mattingly (Duke University)	Redistricting

In addition to the speakers provided by the Census Bureau at the in-briefings, JASON also engaged experts in privacy, demography, and redistricting for virtual briefings and discussions (Table 2-5).

JASON conducted the study from June–September 2021, holding regular team meetings twice a week among the group of JASONs active in the study to coordinate our activities and discuss and reach consensus on our findings and recommendations. Most meetings were held in a hybrid format, with many JASONs participating from our conference room in La Jolla, and others connecting remotely. The JASON members involved in the study were all sworn into Title 13, and some Title 13 materials were provided to us, but the study is not dependant on any restricted materials and all of the contents of this report are unrestricted.

JASON is grateful to everyone who briefed JASON for their important contributions to this study, and to the Census Bureau for its helpful and timely responses to all of our questions over the course of this study.

This Page Intentionally Left Blank

3 USES OF CENSUS DATA

The Census Bureau has a large and diverse set of stakeholders that rely on the Bureau's data products. These are individuals or organizations invested in or affected by the data products produced by the Census Bureau or by outcomes that result from an intermediary that uses the data, as in the distribution of $1.5 trillion in benefits to state and local governments, nonprofits, businesses, and households across the nation (Reamer, 2018; Hotchkiss & Phelan, 2017). Here, we describe some of the users and uses of data products derived from the decennial census. Section 3.5 discusses how those users view consistency as a critical property of census data.

3.1 Stakeholders

The array of stakeholders who use data products released by the Census Bureau is vast and includes a broad range of different types of users. Any list of stakeholders is bound to be incomplete, but we describe several stakeholder communities below to give a sense of the diversity of the Census Bureau's stakeholder communities.

Public Policy Community. This community uses Census data products to support public policy creation, program development, and administration. This stakeholder community exists at the federal level, within federal agencies and federal legislatures, and at the tribal, state, and local government levels. The uses of the Census data products range from the distribution of approximately $1.5 trillion annually of federal funding to the drawing of political districts (which we discuss in more detail in Section 3.2) and the placement of roads, schools, and the provision of emergency services.

Business and Commerce Community. The business community makes extensive use of Census data products through the repackaging of the data to create "value-added" products and services and deriving economic context from the data

through analysis to inform strategic decision-making such as where to locate a new headquarters. The stakeholder community ranges from global corporations to local start-ups.

Non-Government Organizations (NGOs). NGOs use Census data products to support advocacy and to develop and analyze their programs. This community includes special interest groups such as community and cultural associations, religious groups, labor unions, professional associations, and foundations. They may use Census information to assess the socio-economic conditions of a specific group, to evaluate the need to implement various programs, and to monitor the effectiveness of specific programs.

Researchers. The research community uses Census data products as the foundation for a vast array of economic, social, and policy research. Researchers and educators in the social and economic sciences traditionally make up the bulk of this community. Today, however, many other areas including public health, medicine, environmental sciences, engineering, and even the humanities, use Census data products to support their research and deliver educational programs. This community includes both academic and governmental agencies (e.g., the Census Bureau itself, the Bureau of Economic Analysis, USDA Economic Research Service) and other research organizations such as the Brookings Institution, the Urban Institute, and the RAND Corporation.

Journalists. The news media uses Census data products to inform the general public about the socioeconomic status of the residents of the United States at local, state, regional, and federal levels.

General Public. Everyone in the United States is impacted by programs and policies informed by Census products including resource allocation done by their local community and federal financial assistance and tax credit programs.

3.2 Redistricting

An essential use of Census data is for redistricting, which is the process of dividing an area into geographic units for a variety of purposes, including elections and local administration. Election districts include Congressional districts, state legislative districts, county commissions, and city councils. Census data are also used to draw boundaries of school districts, for zoning and planning, and to locate and manage fire and emergency management services.

The disclosure avoidance modernization raised new concerns about the data that will be available to support the range of redistricting needs (Kirkendall et al., 2020), and the requirements for the redistricting use case were an important factor in driving decisions about the privacy loss budget and allocation (as described in Section 5.4). Most of this stems from the noise and biases that are expected in the block-level data. Specific to the PL data are various concerns about states being able to use these data to meet their "one-person one-vote" principle which requires voting districts to be roughy equal in population. Numerous lawsuits have been fought over exactly how this should be interpreted (Rosenberg, 2009). A 1964 Supreme Court decision in *Reynolds v. Sims* disallowed a plan by the state of Alabama to allocate state senators by county despite county populations varying by over a factor of 40. Subsequent court rulings have refined the meaning of this requirement. For example, *AFL-CIO v. Elections Board of Wisconsin* (1982) ruled that an acceptable redistricting plan should have district sizes that deviate by less than 2%. The concept that districts could be exactly equal in size has always ignored the intrinsic error of census counts, and redistricting efforts that try to match populations exactly are succumbing to the *statistical imaginaries* we discuss in Section 7.

Different technical studies in the literature have come to differing and often inconsistent conclusions on the effect of the disclosure avoidance mechanisms on redistricting data (Cohen et al., 2021) (we discuss the Census Bureau's experiments on data reliability for redistricting in Section 6.2). There are political arguments on

both sides regarding the impact of privacy-noise and post-processing adjustments to the PL data. There is also wide use of data available from other sources for developing districts such as the American Community Survey, voting records, and citizen supplied information. These data sources can be used to develop additional dimensions of districts beyond race and ethnicity for districts like "communities of practice" (Wang et al., 2021). Ongoing work is developing mathematical approaches to redistricting (Wang, 2021b; Mattingly, 2021), and these tools could also be applied to bring a better understanding of the impact of privacy noise and data sources on redistricting.

3.3 Demography

Local and state government demographers, business analysts, and researchers, use Census data to study patterns of change in the population. The geographic resolution of these studies needs to be sufficient for the needs of the supported application.

Demographers frequently rely on accurate age pyramids to capture population changes and plan future investments such as the locations of schools and services (French, 2014). Businesses may use this information to decide on new store locations or plan product introductions. This requires a level of demographic detail that does not exist in the PL data, which only distinguishes voting age (18 years and older) population, but is provided in the DHC release. The specific geographic detail needed depends on the application, but most use aggregations of census tracts. For example, in New York City the data are aggregated at the level of boroughs (Salvo et al., 2013) and community districts within boroughs (New York City Department of City Planning, 2021) to support planning and zoning activities. Community districts in New York City cover populations from 50,000 to over 200,000, each district comprising many census tracts.

In other examples, one might want to distinguish between urban and rural places. This is more complicated because the urban and rural geographies are not on the main Census geographic spine. However, it is the aggregated data and not the block level data used to build the aggregates that is essential for most demographic uses.

There are some use cases for decennial data products that benefit from block level data. Several were highlighted in the National Academies of Science, Engineering, and Medicine workshop that explored the implications of privacy noise adjustments to decennial data (Kirkendall et al., 2020). For example, transportation planning benefits from accurate block level counts of both the population and housing units. Such data are available through the released PL data, and does not need the detailed demographic attributes provided in the DHC data.

Geographic areas that are below the level of the state and not coincident with the geographic boundaries on the Census spine are considered small areas. Programs exist to build demographic estimates for these areas (United States Census Bureau, 2021f). These programs frequently build their estimates by modeling census block level data for population and housing unit totals (Mohrman & Kimpel, 2012). These are model-based estimates and they may use ACS or other survey data, but they do not tend to use more complex block level data typically supplied by DHC.

Some recent examples include the ways Census data products are being used to monitor and allocate resources for the Covid-19 pandemic. For example, the Center for Disease Control (CDC) Social Vulnerability Indices (Flanagan et al., 2011; Agency for Toxic Substances and Disease Registry, 2021) have been used to help guide the placement of testing and vaccination sites (Arling et al., 2021; Hughes et al., 2021). These, and similar indices, are built at the Census tract level.

3.4 Tribal Areas

Whereas Census tract or block-group data are suitable for most demographic use cases, Tribal Areas are an exception where detailed block-level geographic data may be needed for important use cases. Tribal Areas are traditionally considered *off-spine* geographies. As such, they are composed of collections of blocks at the bottom of the Census spine that do not form larger on-spine geographic units such as counties (see Figure 2-1 and discussion in Section 5.2). Tribal governments have emphasized to the Census Bureau their need for accurate statistical data at the census block level (Allis, 2019; Gomez, 2021). On sparsely populated tribal lands, census blocks are often geographically large, and the next level up (census block groups) so large as to be almost useless.

Article 1, Section 8, of the U.S. Constitution gives Congress the exclusive power to regulate commerce "with the Indian Tribes". A succession of Supreme Court decisions have established three principles of American Indian law: (1) that tribal authority on tribal lands is organic and not subject to state control; (2) that Congress, not the Executive Branch, has ultimate authority on matters affecting the tribes; and (3) that the federal government has a "fiduciary duty" to protect the tribes (McCarthy, 2004, p. 20).

It seems perfectly consistent with Congress' plenipotentiary role and "duty to protect" that it might legislate different rules for the release of census data for geographic regions that are entirely on tribal lands. The Census Bureau should use its convening powers to explore with tribal governments whether there is a consensus on what these different rules might be. Tribal governments may prefer a different balance of between collective and individual welfare than do correspondingly sized non-tribal communities. Thus, the balance point between Census privacy and accuracy that is politically optimal for the United States may be quite unsatisfactory for certain sovereign tribal nations. One possibility short of the public release of block-level data without privacy noise might be to formally release each tribe's unadjusted data to that tribe's established sovereign government.

3.5 Need for Consistency

Stakeholders are accustomed to receiving the same types of data products that have been historically disseminated, and to viewing an official statistic provided by the Census Bureau as the one, gold standard, number to use as provided. Changes to these products may unsettle users of the data. For example, changes in the way the census forms collect race and ethnicity and how the Census Bureau records this information between the 2010 and 2020 censuses has generated confusion on how to compare the these data across the decades (Hansi Lo Wang and Ruth Talbot, 2021). This is one example of a change motivated by improving data quality leading to inconsistency across time. Such changes require a careful explanation to understand how to effectively use the data and what types of comparisons to avoid, but often the benefits of such a change outweigh the costs to consistency.

Inconsistencies in survey data are frequently attributed to data quality issues in the collection or editing of the data. The U.S. Office of Budget and Management publishes guidelines for statistical surveys to ensure such data quality and consistency (2006). Consistency in measurement is also a concern for data that will be analyzed over time or across different entities such as comparing metrics between countries (National Academies of Sciences, Engineering, and Medicine, 2017). Preserving consistency in these cases is important and leads to more efficient data collection, processing, and use.

Our main focus is on formal consistency properties as defined in Section 2.2. These formal consistency properties ensure that released data is free from internal contradictions. For example, the total population of a state should equal the sum of the populations of all the component counties in the state. Formal inconsistencies can naturally manifest in privacy-protected data since the methods for ensuring privacy involve adding independently-sampled random noise to each value (see Section 4.3). These inconsistencies are not a sign of problems with the quality of the data collection or processing, but a direct result of the privacy noise. Without the extra post-processing steps the Census Bureau performs for the PL data release,

it would be likely that different ways of computing semantically equivalent values would produce different results.

Stakeholders, including several who spoke with JASON, have expressed a strong desire for consistent data across the decennial data products. Such consistency has historically formed the bedrock of statistical use of the data, making it seamless to combine data across tables and easy to integrate it into traditional models and analyses. Arguments for consistency focus on worries about unknown problems stemming from inconsistent data, including how to properly use it. While compelling, the theoretical and the technical reasons for needing such consistency remain unclear. New forms of estimation may need to be developed that can ingest the statistically inconsistent data and account for the uncertainties. But, these can and should be developed on a sound statistical basis.

The most convincing argument JASON could find for consistency is that if there are multiple ways to produce a value users may select which value to use in strategic ways. For example, in scenarios such as funding allocation where a user has a reason to prefer a higher value, they will be tempted to "cherry-pick" among the possible values and select the highest one that can be produced. Others competing for the same resources will search for a way to compute the lowest possible value, leading to unfairness due to the technical abilities of different users to find maximal values, as well as opening the door to political and legal battles over which values to use.

JASON agrees that such a scenario would indeed be chaotic and problematic. But, this could be prevented by the Census Bureau publishing clear and definitive guidance on how to use the data products, including transparent dissemination of calculations that were done to derive any official values. The Census Bureau can also mitigate the problem by providing recommended methods to use for producing statistics from the published data in ways that would make it clear which value should be used when there are multiple ways to produce a value that may result in inconsistent results. This is similar to what the Census Bureau currently does

for other data releases, such as the American Community Survey (United States Census Bureau, 2021b).

This Page Intentionally Left Blank

4 PRIVACY BACKGROUND

There is an inherent trade-off between privacy, which is achieved by either withholding or perturbing data, and utility, which benefits from comprehensive, accurate, and unperturbed data. Perfect privacy is only ensured if no information at all is released, but releasing no information has no utility.

In this section, we provide background on disclosure avoidance as historically used by the Census Bureau, and the evidence from disclosure attacks that moved the Census Bureau to adopt a new approach for the 2020 census. Section 4.3 provides background on the formal privacy notions adopted by the Census Bureau. Section 4.4 describes some other deployments of privacy mechanisms designed to satisfy formal privacy, and how the 2020 census use is different.

4.1 Disclosure Avoidance Methods before 2020

The Census Bureau has been concerned with keeping data confidential going back at least two centuries, with laws requiring confidentiality and providing penalties for disclosures since the 19th century (United States Census Bureau, 2019b). Over the previous five decennial censuses, the Census Bureau adopted a variety of disclosure avoidance methods that evolved to provide stronger protections (McKenna, 2018). For the 1970 census, full tables were suppressed for areas with low population or household counts, but the information in those tables could sometimes be derived from complementary tables that were not suppressed. For the 1990 census, the Census Bureau introduced a data swapping method that would identify households at risk of disclosure and swap them with other households in ways that preserve some properties but vary others in the swapped households, and preserving totals over the area that included both households. These methods were extended for the 2010 census with methods to generate partially synthetic data to protect group quarters from disclosure.

Although the methods used in previous censuses appear to have been sufficient to avoid any major disclosure harms (at least, there are no cases known to JASON in which individuals claimed to be harmed because of individual data disclosed through the 1970–2010 census data releases). However, the traditional disclosure avoidance methods do have drawbacks which prompted the Census Bureau's move to develop modern disclosure avoidance methods for the 2020 census.

The first drawback is that the traditional methods are necessarily *non-transparent*. The Census Bureau has not released its criteria for determining when a household is swapped, or publicly release statistics on the rate. The effectiveness of the swapping protections depends on keeping information about the swapping criteria and rate secret—as a simple example, if an adversary can determine that no households in a particular block were swapped, there is no privacy for individuals in that block and the adversary knows that any inferences they can make about those individuals are valid (at least with respect to the underlying census microdata). The adoption of fully transparent disclosure avoidance mechanisms for the 2020 census, where the only secrets are the random numbers used to sample from the privacy noise distributions, provide tremendous benefits in that the methods can be publicly discussed, analyzed, and their impact on data can be evaluated. The substantial transparency of the methods adopted for the 2020 census benefits the public, but has opened the Census Bureau to a high level of public scrutiny including lawsuits (State of Alabama, 2021).

The second drawback is that the confidentiality protections provided by the traditional methods cannot provide any privacy guarantees. The experiments the Census Bureau did to evaluate the reconstruction and reidentification risks from 2010 census data (described in the next section) raised alarm that the methods used were not providing sufficient protection. JASON evaluated these experiments in our 2019 study, and our conclusions concurred with the concerns raised by the Census Bureau and the need to modernize disclosure avoidance methods for the 2020 census (JASON, 2020).

Another drawback of the traditional methods is the impact they have on data accuracy and possible biases they introduce are unknown and unknowable to users. Because of the need to keep the specifics of the methods secret, the Census Bureau cannot reveal statistics that clarify their impact on data accuracy without compromising the privacy protections. With past censuses, users accepted the released data as "ground truth" even though certain statistics were affected by the swapping and suppression. Users could do nothing to assess the impact of the disclosure avoidance methods on conclusions they might draw from analyzing the released census data.

4.2 Disclosure Attacks

Until recently, the disclosure avoidance methods described in Section 4.1 were considered sufficient to protect the confidentiality of the sensitive microdata from which they are derived. Increases in computing power, and improvements in algorithms, and theoretical results on database reconstruction, however, raised questions about how much an adversary could learn from the published tables. In 2003, Irit Dinur and Kobbi Nissim proved that with sufficiently many queries a database could be reconstructed with near-perfect accuracy even when noise is applied to the responses.

In 2018, the Census Bureau looked at the feasibility that the tabular summaries could be processed to infer the microdata records that were used to produce them (Abowd, 2019). This had not been thought to be feasible owing to the large amount of data and computation involved. According to Census Bureau internal policies, any form of data released that is in the form of microdata needs further review to determine if it satisfies Title 13. Such reconstruction of the microdata does not, by itself, necessarily constitute a violation of Title 13 since the reconstructed microdata contain no personal data (it contains only attributes, not names or addresses), just the individual attribute records in the underlying data used to build the tables. But, as in other re-identification attacks, if external

data can be joined with the microdata then it may be possible to link records in the reconstructed microdata with personal identities available from those external sources. If an adversary can learn new information about sensitive attributes of an individual from this re-identification which is not available in the external sources, such as their race or composition of their household, this would constitute a disclosure of sensitive information due to the census data release which would be interpreted as a violation of Title 13. JASON studied these reconstruction and reidentification attacks in detail in a 2019 study (JASON, 2020), so we only provide a brief summary here.

To test the potential for reconstruction and reidentification attacks on the 2010 census data products, the Census Bureau conducted experiments using a subset of the published data consisting of the nine tables listed in Table 4-6. These tables included detailed block-level data, as well as one table with tract-level detailed age, sex, and race information.

Each cell in each table can be viewed as an integer-valued linear equation over the microdata. For example, if we set the count of people in tract t who are male and who are 27 years old to $T_{t,M,27}$ then this is tabulated via the equation $\sum_p B_{t,M,27}(p)$, where p sums over the internal identification number and $B_{t,M,27}$ represents a indicator function that is 1 if the record corresponding to p has attributes matching a block in tract t, sex male, and age 27 and zero otherwise (Leclerc, 2019). To solve the resulting set of equations, the Census Bureau used a commercially-available optimization solver, Gurobi (Gurobi Optimization, LLC, 2019). The Gurobi solver attempts to find an integer solution to the set of equations corresponding to the tabulations. To break up the problem into manageable pieces Census applied the solver at the tract level. The solver was able to solve the resulting systems of equations to produce microdata for the entire U.S. covering all 70,000 Census tracts and all 6.4 million potentially populated blocks from the 2010 census. The resulting reconstructed microdata contained a record for each individual in the reconstructed dataset, with the following information: their census block, a binary attribute indicating if they are of Hispanic origin (or not), a race attribute

Table 4-6: Tables used in 2018 Census Bureau reconstruction experiments on 2010 census data. (The 2020 table identifiers are based on the mapping in the planning crosswalk (United States Census Bureau, 2020a). Note that because of the data consistency and semantics of these tables, tables P1, P6, P7 and P9 could be derived from the information in table P11, and table P12 could be derived from the P12A–I tables. The plans for the 2020 DHC do not include any tables corresponding to the 2010 PCT12A–O tables at the tract level, but do include similar information at the county level.)

Table (2010)	Table (2020)	Granularity	Contents
P1	DHC-P1	Block	Total population
P6	DHC-P6	Block	Total races
P7	–	Block	Hispanic or Latino origin by race
P9	PL-P2	Block	Hispanic or Latino and not Hispanic or Latino by race
P11	PL-P4	Block	Hispanic or Latino and not Hispanic or Latino by race by voting age ($\geq$ 18)
P12	DHC-P7	Block	Sex by age
P12A–I	DHC-P7A-I	Block	Sex by age, iterated by race (in separate sub-tables)
P14	DHC-P9	Block	Sex by single-year age ($<$ 20)
PCT12A–O	–	Tract	Sex by detailed age, iterated by major race alone

identifying one of the 63 OMB race categories, a binary attribute representing their sex, and a non-negative integer indicating their age in years.

The next step was to see if the reconstructed microdata could be linked with data that would be readily available to a potential adversary. Some data on individuals is available in the public domain, including data that could be assembled using public records. More complete and current data can be licensed through private companies such as marketing research firms. Available public or commercial data may contain the name, address, sex and birth date of each individual, but typically does not contain information regarding race and ethnicity. Using the reconstructed database and auxiliary data (including public and commercial data), the Census

Bureau performed a database join by matching the addresses in the commercial databases to blocks in the reconstructed records, and the age and sex attributes, to produce a joined table where the reconstructed race and ethnicity attributes in the reconstructed records are now connected to individual records in the auxiliary data. When the reidentification is correct, it simulates an attack where an adversary is able to learn sensitive attributes of individuals by combining the released census data with available auxiliary data.

To evaluate the simulated attack, the Census Bureau compared the reconstructed records with the microdata in both the 2010 Hundred-percent Detailed File (HDF), which is the microdata from which the 2010 census data products were derived, and the 2010 Census Edited File (CEF). The HDF is the result of applying the 2010 disclosure avoidance mechanisms (swapping) to the CEF. The Census Bureau determined that 48.34% of the reconstructed records matched exactly to records in the HDF microdata. If a fuzzy match on age were used to allow ages to match if they are within one year, 73.33% of the records matched. In comparing the reconstructed records to the original 2010 CEF data, the Census Bureau found 46.48% (138 million) matched exactly with the CEF records, and 70.98% matched within one year of age. Of the 138M matched records, 37% corresponded to correct reidentifications indicating that 52 million people (17% of the enumerated population) could have been reidentified.

One limitation of this attack experiment is that the success rate depends on the auxiliary information available to the adversary. These simulated attacks assumed the adversary only had access to a small amount of readily available data. To estimate the worst case for auxiliary knowledge, but not the worst case adversary who may employ more sophisticated attacks, the Census Bureau conducted additional experiments assuming an adversary had the best possible auxiliary data. In this case, the best possible data would be the original CEF data, but without the race and ethnicity information. These data were then used to see if a simulated reconstruction attack on the released 2010 data products could infer the removed values. The attack found 238 million (77%) putative reidentifications,

of which 75% (179 million), covering approximately 58% of the population, were found to be correct reidentifications when checked against the original microdata.

4.3 Formal Privacy

The disclosure risks revealed by the Census Bureau's reconstruction experiments, and the other drawbacks associated with the traditional disclosure avoidance methods, led the Census Bureau to adopt formal privacy notions that had emerged from the computer science community over the previous two decades for the 2020 decennial census data products. These notions give formal and quantifiable guarantees on inference disclosure risk and known algorithmic mechanisms for releasing data that satisfy these guarantees. Many detailed references on those notions are available (e.g., Dwork & Roth 2014), and we do not attempt to provide details on the formal results here. Instead, this section aims to provide enough background to understand what the notions used by the Census Bureau mean and their most important implications for the 2020 census.

4.3.1 Differential privacy

Differential privacy was introduced in a 2006 paper by Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Since then, differential privacy, and variations on the original differential privacy definition, have become the dominant formal privacy definitions. The original definition, which we refer to as *pure* differential privacy to distinguish it from the relaxed differential privacy notions introduced later, gives a bound on the inference risks associated with data releases produced by a randomized algorithm:

Definition 4.1 (Pure Differential Privacy)**.** A randomized algorithm $\mathcal{M}$ is (ϵ)-*differentially private* if for any pair of neighboring data sets S, S', and any set of outputs O,

$$Pr[\mathcal{M}(S) \in O] \leq e^{\epsilon} Pr[\mathcal{M}(S') \in O].$$

The privacy loss budget parameter, ϵ, provides a way to control the privacy-utility tradeoff in a quantifiable way: higher values of ϵ result in an exponentially higher inference bound and correspondingly lower privacy guarantee.

The definition is very strong—it provides an information-theoretic guarantee that requires no assumptions about the methods and resources available to an adversary attempting to make inferences from the released output. In particular, unlike many earlier attempts to develop formal privacy notions, it does not place any limits on the information the adversary may already know—the inference bound holds regardless of any information sources available to the adversary. It also does not depend on the actual data—the inference bound holds for any data set S, and any of its neighbors, S'. What it means for a data set to be neighboring is flexible, but the choice of what neighboring means has implications for the meaning of the privacy guarantee and the mechanism that must be used to satisfy it. For databases, a typical definition of neighboring is differing by one record. Since each record corresponds to the microdata for one individual, this means the inference bound limits the disclosure risk to an individual by distinguishing the inference probabilities in the scenarios where their record is, and is not, included in the data set. For the Top-Down Algorithm, the Census Bureau defines neighboring as a single modification to an individual record. For example, changing the race attribute or the Census block for an individual would result in a neighboring dataset.

The differential privacy inference bound also holds for any post-processing done on the data—post-processing could improve privacy, but (so long as it does not use the original sensitive data in any way) cannot compromise privacy. In this sense, the definition provides an inference bound that captures the worst case disclosure risk to an individual of having their data included in the census.

Note that the definition depends critically on the mechanism $\mathcal{M}$ that is used to produce outputs from the source data being *randomized*. To satisfy the definition, the output of the mechanism must employ cryptographically strong randomness

which is unpredictable and kept secret. The output produced for any given input is sampled from a set of possible outputs in a way that depends on that randomness. Satisfying pure differential privacy requires that the sets of possible outputs from any pair of neighboring data sets must be identical. Possible outputs are values for which the probability of output is non-zero, and the inference bound given in the definition must hold for any set of outputs O. If there is any output with non-zero probability for $\mathcal{M}(S)$ but zero probability for $\mathcal{M}(S')$, the mechanism cannot satisfy the pure differential privacy definition. This implies that the randomization use in $\mathcal{M}$, in a general case, must involve unbounded noise.

Composition. Differential privacy satisfies a simple composition property that allows the output of multiple mechanisms to be performed on the same data with a linear composition of the privacy loss budget. When two mechanisms with privacy loss budgets ϵ_1 and ϵ_2 are performed on the same data, together the total privacy loss budget consumed is $\epsilon_1 + \epsilon_2$. This means that a data curator can decide on a global privacy loss budget for their data and then partition that privacy loss budget across all of the mechanisms that will be applied to produce outputs from their data. For a setting like the Census Bureau's use of the decennial census microdata, one can think of each query being done on that microdata to produce cells in a table as one mechanism so that the privacy loss budget allocation can be partitioned across all of the queries to be done.

Differential privacy also satisfies a *parallel composition* property. When multiple mechanisms are performed, each of which touches only a non-overlapping partition of the dataset, the total privacy loss of the set of outputs produced is the maximum of the privacy loss budget for any one of the queries. Thus, when the same query is performed on many geographic units (say for all counties in the United States), the total privacy loss incurred is the privacy loss budget of one of

the queries since each query touches a different partition of the full microdata.[1]

Laplacian Mechanism. One way to achieve ϵ-DP is to add noise sampled from a Laplace distribution (Dwork & Roth, 2014). The Laplace distribution is unbounded. The scale of the noise must be proportional to the *sensitivity* of the underlying function, Δ, which is the maximum change in the output over all possible pairs of neighboring datasets. So, if $f(D)$ is the original algorithm, the randomized mechanism $\mathcal{M}(D) = f(D) + \text{Lap}(\Delta/\epsilon)$ satisfies ϵ-DP where $\text{Lap}(S)$ samples Laplacian noise with center 0 and scale S.

Since the definition of *neighboring* used by the Census Bureau is modifying a single record, the L1 sensitivity of all outputs that are counts of individuals is 2, since changing the value of an attribute reduces one total count by 1, and increases another count by 1. The L1 norm is the sum of the differences in the counts between the two neighboring datasets. The L2 (Euclidean distance) sensitivity can also be used, which results in a sensitivity of $\sqrt{2}$, as is used in the Top-Down Algorithm.

Discrete Noise. The formal results on differential privacy and the Laplacian mechanism are proven using real numbers, which cannot be accurately represented on finite computers. Indeed, the least significant bits of revealed outputs could allow pre-noised values to be inferred when floating point approximations of the Laplacian distribution is used carelessly to implement the Laplacian mechanism for differential privacy (Mironov, 2012). To overcome these problems, the discrete Laplacian distribution (also known as two-sided geometric) is used instead. The discrete Laplacian distribution is given by the probability distribution $\frac{e^{\epsilon}-1}{e^{\epsilon}+1} \cdot e^{-|n\epsilon|}$

[1]The actual situation for the Census is more complicated, because the Census Bureau uses a bounded neighboring notion (modifying a record) instead of unbounded (adding or deleting). A single change in the data can affect more than one group, and because the optimized spine causes different geographic units to have unequal noise scales for the same queries. These complexities are addressed in the privacy proofs for the Top-Down Algorithm (Abowd et al., 2021).

over the integers, $n \in \mathbb{Z}$. Adding sampled noise to integer-valued results from an appropriately scaled discrete Laplacian distribution also provides pure ϵ-DP (Ghosh et al., 2012).

4.3.2 Zero-concentrated differential privacy

The initial mechanisms proposed by the Census Bureau and used for the early demonstration privacy-preserving data releases used pure differential privacy with the discrete Laplacian mechanism. To improve accuracy, the Census Bureau later switched to a different formal privacy notion which relaxes the pure differential privacy definition. This change allows noise to be sampled from a discrete Gaussian distribution, which has lower probability for values far from the center than those for the Laplacian distribution that would be used to satisfy pure differential privacy.

The formal privacy definition used by the Census Bureau is *zero-concentrated differential privacy* (zCDP), introduced by Bun & Steinke (2016), building upon an earlier notion of Concentrated Differential Privacy (Dwork & Rothblum, 2016). Zero-concentrated differential privacy is defined in terms of the Rényi divergence between the output distributions:[2]

Definition 4.2 (Zero-Concentrated Differential Privacy)**.** A randomized mechanism $\mathcal{M}$ is (ρ)-zero-concentrated differentially private if, for all neighboring data sets D and D' and all $\alpha \in (1, \infty)$,

$$\mathcal{D}_\alpha(\mathcal{M}(D) \,||\, \mathcal{M}(D')) \leq \rho\alpha$$

where $\mathcal{D}_\alpha(\mathcal{M}(D) \,||\, \mathcal{M}(D'))$ is the α-Rényi divergence between the distribution of $\mathcal{M}(D)$ and the distribution of $\mathcal{M}(D')$.

[2]The original definition includes a ξ parameter, which we set to 0 instead of including here.

The zCDP definition does not correspond as simply to an intuitive notion of privacy as the original differential privacy definition, but there is a connection between ρ-zCDP and the privacy loss budget for differential privacy. If a mechanism $\mathcal{M}$ satisfies ϵ-DP, then it also satisfies ρ-zCDP where $\rho = \frac{1}{2}\epsilon^2$. In the other direction, a mechanism $\mathcal{M}$ that satisfies ρ-zCDP cannot be guaranteed to satisfy ϵ-DP for any (finite) ϵ, but can be converted into an approximate differential privacy notion:

Definition 4.3 (Approximate Differential Privacy)**.** A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-*differentially private* if for any pair of neighboring data sets S, S' that differ by one record, and any set of outputs O,

$$Pr[\mathcal{M}(S) \in O] \leq e^{\epsilon} Pr[\mathcal{M}(S') \in O] + \delta.$$

When $\delta = 0$, this definition is equivalent to pure differential privacy. For non-zero δ, a mechanism $\mathcal{M}$ that satisfies (ϵ, δ)-DP is guaranteed to satisfy ϵ-DP with probability that is at least $1 - \delta$. The value of δ is typically bounded by the reciprocal of the size of the data set. If a mechanism $\mathcal{M}$ provides ρ-zCDP, it satisfies (ϵ, δ)-DP for any $\delta > 0$ where $\epsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$. This means for a given privacy loss budget parameter ρ for zero-concentrated differential privacy, a corresponding value of ϵ can be computed for any selected value of δ.[3]

Gaussian Mechanism. Whereas Laplacian noise was needed to satisfy pure differential privacy, zero-concentrated differential privacy can be achieved by sampling noise from a Gaussian distribution. Adding noise randomly sampled from a Gaussian distribution with variance $\Delta/(2 \cdot \rho)$ is sufficient to achieve ρ-zCDP. The considered advantage of switching to the zCDP notion is that it provides sublinear composition to enable higher utility to be achieved with less accuracy loss.

[3] A slightly tighter, but more complex, conversion is proven in Bun & Steinke (2016), which allows conversions that produce a lower value of ϵ. These are the conversions that were used to produce the published (ϵ, δ) values in the Census Bureau privacy parameter releases.

Similarly to the issues with the continuous Laplacian mentioned earlier, the continuous Gaussian distribution cannot be represented by finite computers. The discrete Gaussian distribution for scale σ and centered around 0 is a similar probability distribution on the integers which can be computed efficiently by finite computers:

$$Pr_{X \leftarrow \mathcal{N}_{\mathbb{Z}}(\sigma^2)}[X = x] = \frac{e^{-x^2/2\sigma^2}}{\sum_{y \in \mathbb{Z}} e^{-y^2/2\sigma^2}}$$

This discrete Gaussian distribution is the noise distribution used to produce the 2020 census data products covered by this study. The Census Bureau constructed proofs showing that the way the discrete Gaussian is used in the Top-Down Algorithm guarantees the ρ-zCDP property is satisfied.

4.4 Deployments of Differential Privacy

Differential privacy has been a tremendous success, both in the research community, and in many high profile deployed applications. Here, we describe a few applications using differential privacy, and contrast these uses to the challenges faces in deploying differential privacy for the 2020 census.

OnTheMap. The Census Bureau has been using noise-based privacy methods since before the invention of differential privacy (Abowd et al., 2005), and was a pioneer in using differential privacy. The Census Bureau used a form of synthetic noisy data for the OnTheMap application by 2006 (Andersson, 2007). Shortly after the invention of differential privacy, a version of OnTheMap was released using differential privacy (Machanavajjhala et al., 2008). This was the first significant production application designed around satisfying differential privacy guarantees.

Industrial Applications. Differential privacy has seen widespread adoption in industry, with applications deployed by Apple (Greenberg, 2016; Apple Computer, Inc., 2021), Google (Erlingsson et al., 2014), and Uber (Greenberg, 2017; Near, 2018), and a number of well-funded companies developing differential privacy

solutions (e.g., LeapYear, Inc. (2021), Oasis Labs (2021), Privatar Labs (2021), and Tumult Labs (2021), which works with the Census Bureau).

The most prominent industrial applications, such as Apple's use of differential privacy to collect information about user's use of emoji and words in confidential text messages, and Google's use of differential privacy to collect statistics about browser crashes, provide what is known as *local differential privacy*. With local differential privacy, the privacy notion and protection methods are the same as when differential privacy is used to protect aggregated statistics, but privacy protections are applied to individual data records before they are collected and analyzed by the central server. This provides desirable privacy protections to end users since their confidential data never leaves their own control without privacy noise already having been applied to protect it. Local differential privacy can work in settings where the goal is to produce a limited number of aggregate statistics from a large population of users, such as Apple's use to collect information about popular emojis. Even in such applications, however, there are questions about how much real privacy is provided. The privacy loss budget selected is fairly high ($\epsilon = 4$ for the emoji application) and is reset every day for each user. Further, because of the proprietary nature of the implementation, a user is left to rely on promises made by the vendor that privacy properties are indeed satisfied. This can be perilous—an analysis of Apple's implementation found numerous flaws that prevented it from satisfying stated privacy claims (Tang et al., 2017).

Comparison to the Decennial Census Data Products. Achieving differential privacy for the 2020 census data products poses numerous challenges that are unlike those faced by industrial applications. Most industrial uses of differential privacy use local differential privacy, applying noise to data before it is collected in a centralized database. Industrial uses of centralized differential privacy such as Uber's are often primarily focused on limiting corporate liability from data abuses by a rogue employee. In these cases, the data are already fully available to the company and are used for internal analysis, not for producing public data products.

The Census Bureau has an obligation to protect the confidential data entrusted to it and to release high-value public-use data products derived from that data. Unlike typical industrial applications which release few, if any, statistics from a huge amount of privacy-noised data, the Census Bureau needs to release billions of statistics from a relatively small amount of data: one microdata record on each person in the United States, with just a handful of attributes for that record.

For all the industrial applications, individuals have some choice whether or not to provide their data to the company. In some cases, such as the browser statistics Google collects, there is an explicit option to opt-out of data collection. In other cases, individuals at least have the option to avoid the company's products. This aligns well with the intuitive notion of differential privacy as a bound on the risk to an individual of contributing their data to a data set. Individuals do not have the same option with the decennial census. People are required by law to participate, and the Census Bureau is required to do a full and accurate enumeration. An individual who declines to self-report is still likely be included in the census microdata from proxy interviews by neighbors or through administrative records such as federal tax information provided by the Internal Revenue Service.

Importantly, the confidentiality requirements on the Census Bureau are different from those faced by industry. Title 13 applies to the Census Bureau and imposes severe penalties for inappropriate data disclosure. Beyond the legal requirements of Title 13, the Census Bureau's long term effectiveness depends on preserving its reputation as a trusted data collector and curator for the American people.

An additional complication for the Census Bureau use of differential privacy mechanisms is the internal requirement to generate synthetic privacy-protected microdata. Instead of just releasing the privacy-protected statistics resulting from the differentially private mechanism applied to the original sensitive microdata (as is done in all industrial uses of which we are aware), the Census Bureau imposed an internal requirement that the data products must be produced from synthetic

microdata generated to produce the same results as the privacy-protected statistics when the queries are run on that data. This requirement was a purely internal one, and there are no expectations that the synthetic microdata will be released outside the Census Bureau, but it presented many additional challenges to the design of the disclosure avoidance system.

Finally, the decennial census happens once every ten years under intense political pressure with only one chance to get it right (Bazelon & Wines, 2021). This is very different from industrial applications where an application can incrementally improve over time as it scales to support a larger user base. The Census Bureau does not get the opportunity for trial-and-error for a decennial census that is afforded to most organizations.

5 DISCLOSURE AVOIDANCE FOR THE 2020 CENSUS

As discussed in Section 4.4, the Census Bureau faced a number of challenges in implementing disclosure avoidance mechanisms that provide formal privacy guarantees for the 2020 census data products. Over the years leading up to the 2020 census, the Census Bureau developed a plan to modernize its disclosure avoidance methods based on the differential privacy notion. This planning was done through a transparent and open process with stakeholders engaged throughout the process. The eventual methods used to provide privacy in the 2020 census data products evolved in response to internal Census Bureau studies, developments in the academic research community, and feedback from stakeholders.

Figure 5-1 depicts the process the Census Bureau uses to produce privacy-preserving data products from the CEF microdata. A few statistics are *invariants* that are released without any privacy-protecting noise. State population totals are invariant. Although the Constitution requires that apportionment be done based on the "actual enumeration" of the population, it does not necessarily require that the actual counts be disclosed without perturbation. However, guaranteeing invariance in representation without making the state population counts invariant would be

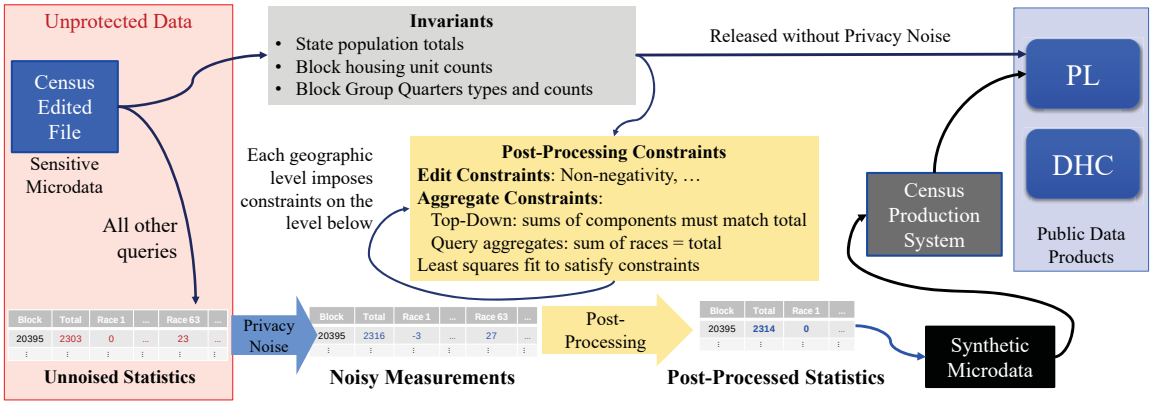

Figure 5-1: Process used to produce privacy-protected data products.

complicated and politically fraught, and the Census Bureau's Data Stewardship Executive Policy Committee determined that the disclosure risk associated with releasing these counts without privacy noise was acceptable. The counts of block housing units and occupied group quarters by type are also invariant because they are considered public information. All other statistics derived from the CEF are only released after privacy protections are applied.

The process used to produce these statistics is to first execute queries on the sensitive microdata in the Census Edited File. The results of those queries, the *unnoised statistics* in the figure, are still sensitive. Then, privacy noise is added to those values. This produces what is known as the *noisy measurements*. This data is now privacy-protected. Any further adjustments to this data cannot reduce the privacy properties established by the privacy noise (so long as they do not involve additional access to the sensitive microdata).

Next, the noisy measurements are adjusted by post-processing to produce the post-processed statistics. The post-processing done on the noisy measurements adjusts the values to produce data that satisfies certain constraints, such as satisfying formal consistency properties and avoiding negative counts. Synthetic microdata is generated that corresponds to the adjusted data. The synthetic microdata are in the same format as the original CEF microdata, consisting of one synthetic record for each enumerated person. The synthetic microdata are generated such that queries executed on them produce exactly the same statistics as those in the post-processed data. Because of the post-processing ensures formal consistency of the resulting tables, there must exist at least one set of synthetic microdata that has this property. This synthetic microdata is the input to the Census production system which produces the released data products.

In the following subsections, we provide more detail on the algorithm used, how the Census Bureau optimized the geographic spine to improve the results, complexities involved in producing synthetic microdata, and how the production parameters were determined for the August 2021 PL data release. We defer

evaluation of the impact of the disclosure avoidance mechanisms on the census data products to Section 6.

5.1 Top-Down Algorithm

The 2020 census data products include tables at multiple geographic granularities (Table 2-2), from the full nation down to the over 5 million individual census blocks. A natural way to produce these statistics would be to first compute all the statistics at the lowest geographic unit (block), adding privacy noise as necessary to these statistics, and then producing the statistics for larger geographic units by just summing up the corresponding statistics for their constituent components. Such a bottom-up (block-by-block) approach is simple, easy to understand, and would guarantee formal consistency, since the statistics for larger geographic units are produced by summing their component statistics.

However, there would be some serious drawbacks of this method. First, state population counts are required to be reported as enumerated, without any privacy noise. These population counts are mandated by the United States Constitution and are used to apportion representatives in Congress. Small differences in these numbers can be enough to change which state receives a representative. A margin of 26 people in the 2020 enumeration gave the last seat going to Minnesota instead of New York (Wang, 2021a)). The state totals computed using a bottom-up algorithm with privacy noise would not be expected to match these enumerated values, and there is no obvious way to impose such a constraint on a bottom-up algorithm.

The second, and more fundamental, problem with a bottom-up approach is that it provides no way to control the error at higher levels in the geography. The expected error at each level accumulates the errors of its sub-components. The methods the Census Bureau uses to satisfy differential privacy require that noise is sampled randomly and applied independently to each measured value. This would

result in estimates at the block group and higher levels that have expected squared error at least as large as the sum of the variances of its component blocks.

The Census Bureau's solution to these problems is to compute statistics for each geographic level in a top-down manner. This allows the Census Bureau to partition the overall privacy loss budget given to a set of statistics across the geographic levels and across the queries needed to produce the statistics at each level in a way that allows for control of the variance for each statistic.

The top-down algorithm used by the Census Bureau to produce the PL and DHC data products from the decennial microdata can be viewed as having two phases. In the first phase, noisy measurements are taken at each geographic level for each statistic. In the second phase, adjustments are made to these noisy measurements to produce the published statistics at each level, going level-by-level starting at the highest (nation) down to the lowest (block) level. It can also be viewed as a single phase algorithm, proceeding down the geographic levels. At each geographic level, the queries are done to obtain the statistics, and then privacy noise from the distribution determined by the privacy loss budget for that query at that geographic level is added.

Next, the values are adjusted to satisfy constraints. The constraints that are to be satisfied include:

- *Invariants* — certain statistics are required to be published directly as enumerated without any perturbation: state population totals and by-block housing unit counts and group quarters types and counts. Any privacy-noised queries that correspond to these results are replaced with their enumerated values, and the state population totals are treated as any other aggregate constraint at the next level (so the sum of the county populations in a state much equal the invariant state population total).
- *Edit constraints* — certain values that are considered impossible are not permitted in output tables. For example, the population count must be at

least one for each occupied group quarters facility of each type. The full set of edit constraints used by the Census Bureau is not made public, but there are only a few applied for the PL data product (Abowd et al., 2021). More complicated edit constraints apply to other data products that involve queries that require joining the persons and housing units files, such as a requirement that the number of households cannot be less than the number of householders since each household must include at least one person.

- *Non-negativity constraints* — values that are counts of people cannot be negative. Technically this is an edit constraint. Because of its importance it is given a distinguishing name.

- *Aggregate constraints* — the sum of all the component statistics must total the total statistic. This includes geographic aggregates where the sum of any statistic over its component geographies should equal the total for the aggregate unit. For example, the sum of the values of that statistic for all counties in the state must equal the state total, and so on down the geographic hierarchy. Attribute aggregates must also satisfy aggregate constraints. This means the sum of a statistic over a set of attribute values that partitions the whole population must equal the total for that population. For example, the sum of the male and female voting age populations for a block must equal the total voting age population for that block.

- *Non-fractionality constraints* — all values must be integers.

The types of constraints have different names due to their sources. All the constraints, except for the non-fractionality one, can be expressed mathematically as sets of statistics that must either sum to an exact equality or be greater or equal to a particular total. Specifically, the aggregate constraints require the sum of a set of component statistics to equal the total statistic, the non-negativity constraints require that all statistics to have values ≥ 0, and the requirement that occupied group quarters have positive population is a ≥ 1 requirement on particular statistics.

The Top-Down Algorithm adjusts the values resulting from the noisy measurements at each level to satisfy the desired constraints. First, all the constraints other than non-fractionality are satisfied by finding an adjustment that minimizes the sum of the squares of the induced errors. Then, those values are adjusted to integers to satisfy the non-fractionality constraint by using a controlled rounding process that preserves the necessary sum equalities.

The Census Bureau developed several improvements to this algorithm that led to better utility with the same privacy loss budget. These involve dividing the constraints into multiple passes solved in orders that minimize the distortions caused by the non-negativity constraint and by combining the controlled rounding and constraint solving steps (Abowd et al., 2021).

5.2 Optimizing the Geographic Spine

The Top-Down Algorithm allocates privacy loss budget at each level in the census geographical hierarchy in Figure 2-1. This enables more control over the accuracy of statistics for the entities on the geographic spine than statistics computed by composing areas for estimating other regions off the spine. To generate statistics for a particular region of interest published statistics from geographic units on the spine can be combined, using both addition and subtraction, to find a set of units that corresponds to the desired region. For any region that does not sub-divide census blocks there is always some way to compute a statistic for the region by combining statistics from on-spine units. This is because the spine includes all the blocks.

The accuracy of off-spine geographic regions is affected by the number of on-spine units that need to be combined to create the statistics. For example, if an off-spine geography intersects multiple states and several counties or tracts, this accuracy will be degraded as compared to a geography that is nested within the geographic units on the spine. This can be captured in a mathematical notion of

off-spine entity distance, which is the minimum number of on-spine geographic units that must be added or subtracted together to derive the off-spine entity.

Politically important areas may have large off-spine entity distances, and the problem is especially severe for tribal areas which are often far off-spine and have small populations. To mitigate the impact of privacy noise on these entities, the Census Bureau developed an alternate geographic spine to improve the statistics for some off-spine entities.

5.2.1 American Indian and Alaskan Native areas spine

Geographic regions corresponding to American Indian and Alaskan Native (AIAN) areas were often far off the standard spine. For example, the Navajo Nation in Figure 5-2 spans discontiguous areas in New Mexico, Arizona, and Utah. Applying the 2020 disclosure avoidance methods on the standard spine presents a high risk that some AIAN areas of would have unacceptable error. Advocacy groups for American Indians expressed serious concerns about both the loss and accuracy of important data to them in briefings to JASON (Roubideaux, 2021), in a series of letters to the Census Bureau JASON (Allis, 2019; Gomez, 2021), as well as in a formal resolution requesting consultation on privacy and accuracy (National Congress of American Indians, 2019).

To improve the accuracy of census data for AIAN areas, the Census Bureau created a second geographic spine to directly include these areas, known as the *AIAN spine*. For the 37 states that contain any AIAN areas, an additional state-level geographic unit is created that represents all the AIAN areas in that state. Then, the AIAN areas may be divided into other areas such as State Designated Tribal Statistical Areas, and ultimately into Census blocks. At the state level, the total population counts (which are invariant) include the population in the state's AIAN areas. The state-level AIAN populations are privacy-noised and privacy loss budget is allocated to these statistics. At the block level, every block

Figure 5-2: Map of Navajo Nation (this includes the Navajo–Hopi Joint Use Area). Red dots represent concentrations of Navajo population. (Source: https://en.wikipedia.org/wiki/File:NavajoNation_map_en.svg.)

is either in the standard spine or the AIAN spine with no overlaps. Hence, the parallel composition property of differential privacy applies. The same privacy loss budget is applied to queries at all geographic levels down to the block level, regardless of whether they are in the standard or AIAN spine.

5.2.2 Optimizing groups and queries

To improve accuracy for regions of interest, the Census Bureau developed an algorithm for optimizing the geographic spine to minimize the off-spine distance of important entities including the AIAN tribal areas, as well as of legal and political areas such as townships and incorporated places. The census blocks are fixed as pre-defined, but the block groups that are the level above in the hierarchy are adjustable. The main purpose of the algorithm is to find *optimized block groups* to use in computing the noisy tables. In addition to the block groups, the optimized spine may introduce tract groups when the number of child component

tracts for a county is too large since high fanout causes computational difficulty for the Top-Down Algorithm (but this was found not to be necessary in producing the PL data release). These tract groups can also be optimized to minimize off-spine distances. The details of the algorithm are described in *Alternative Geographic Spines* (United States Census Bureau DAS Science Team, 2021). The algorithm outputs a redefined spine that reallocates blocks to block groups and tracts to tract groups to minimize the off-spine distance of a defined set of geographical areas.

As a further optimization, it is sometimes better for accuracy to bypass a geographic level along some path down the spine. This occurs when the expected squared error of computing the totals at that level by summing up the components at the lower level is less than the expected squared error of the higher-level query. This was done only when the expected square error at the higher level would not increase for *any* of the queries done for this geographic unit (United States Census Bureau DAS Science Team, 2021).

The Census Bureau decided to use the optimized spine only internally, and to release data for the original spine geographic units. The block groups used in data products are known as the *tabulation block groups*. They do not correspond to the *optimized block groups* used to produce the data, but their statistics can be produced by aggregating the containing blocks, just as can be done for any arbitrary collection of blocks. JASON believes it would serve the stakeholders better to release data tables using the optimized block groups. These are the units with the most accurate statistics, and except when historical continuity is required, users should be encouraged to use the optimized block groups.

5.3 Producing Synthetic Microdata

As depicted in Figure 5-1, the post-processed data are not used directly to produce the census data products. It was taken as a requirement for the DAS that it output privacy-protected microdata that could be used directly by the Census Production

System to produce the many different representations of the published tables the Census Bureau publishes for data users. The generated synthetic microdata is known as the Microdata Detail File (MDF) and has the same form as the original sensitive microdata in the CEF with one record corresponding to each individual. This requirement for producing microdata is a result of both technical aspects of the Census Production System and expectations within the Census Bureau that it should always be possible to internally examine the microdata from which a published statistic is derived. The examination of the microdata is the traditional approach for investigating quality issues. The Census Production System was outside the scope of this study so JASON is not able to make a specific recommendation on the decision to produce synthetic microdata. JASON did not examine how difficult it would be to relax this requirement, but understood it as an inviolable requirement for the DAS. We do note, however, that it is an internal requirement—the synthetic microdata are not a public data product and constraining the DAS to produce synthetic microdata has significant drawbacks. The complexities and inaccuracies resulting from this internal requirement are difficult to justify to external stakeholders.

The microdata requirement imposes several constraints on the DAS. Many of the data constraints discussed in Section 5.1 are intrinsically desirable for providing both formal consistency and plausibility properties to data users. But, enforcing them has costs in decreasing the accuracy and introducing biases in the post-processed data. These costs are discussed in Section 6.3.2. If it were not for the microdata requirement, it would be a design option for the DAS to decide which constraints to satisfy based on their expected cost-benefit tradeoffs, and to consider other options such as satisfying constraints for some statistics but not others. The need to produce synthetic microdata, however, limited the Census Bureau's options in relaxing the constraints in two important ways:

1. Every record in the synthetic microdata must correspond to a whole person. This implies both the non-negativity constraints (there is no way to represent

a negative person in the microdata) and the non-fractionality constraints (there is no way to represent a fractional person in the microdata). Although it may be possible to extend the microdata format to represent both negative and fractional persons, it is unclear to JASON how feasible this would be technically for the Census Production System and how acceptable it would be culturally to have a notion of a negative person in the microdata.

2. In order to be able to generate synthetic microdata, the post-processed data must be formally consistent. The requirement for the synthetic microdata is that any query on that data will produce the exact value of the corresponding statistic as in the post-processed data. Hence, it is not possible to produce synthetic microdata that corresponds to inconsistent tables.

The microdata requirement imposes the aggregate constraints, non-negativity constraints, and non-fractionality constraints that are ensured by the top-down algorithm. Without eliminating the internal microdata requirement or extending the microdata format, there is no option to relax these constraints.

The need to produce microdata imposes another detriment to the data production process: the need to perform detailed queries that do not correspond to statistics in the published data products. Although the detailed queries may in some cases be useful in producing more accurate values for published statistics, the need for the detailed queries is often justified by the Census Bureau based on the need to generate synthetic microdata. The consistency constraints ensure that a corresponding microdata must exist, but actually finding one may be a computationally challenging problem. To simplify the process for producing the microdata, the Census Bureau included detailed queries that produce tables with 2,016 cells corresponding to all combinations of the attributes for each geographic unit: Household or Group Quarters Type (8 values) $\times$ Voting Age (2) $\times$ Hispanic/Latino origin (2) $\times$ Race (63). These detailed queries do not correspond to publicly released statistics. These queries necessarily consume privacy loss budget since they use the sensitive CEF microdata and may impact the values

of published statistics. The resulting noisy measurements for these values are then post-processed to satisfy the constraints and used to produce the synthetic microdata.

JASON finds it hard to justify the use of detailed queries to produce synthetic microdata. They consume privacy loss budget to produce statistics that are only needed for internal production purposes and are never released. If the total privacy loss budget allocated to the PL data release is considered fixed, this privacy loss budget could otherwise be applied to other queries to improve accuracy of the released data. If the privacy loss budget is not fixed, eliminating unnecessary queries would reduce the overall privacy loss budget. There are cases where additional queries can be used to improve the accuracy of published statistics, and it may be the case that some of the detailed queries are justified for this purpose. JASON did not find any evidence, however, that could justify all of the detailed queries that are included in the Top-Down Algorithm separate from their use in producing synthetic microdata.

Even if the microdata requirement cannot be relaxed, it is not clear to JASON that it is infeasible to produce satisfactory synthetic microdata without the detailed queries. It would be computationally more challenging to produce microdata from just the post-processed statistics that correspond to statistics in the public data products, but not necessarily intractable. The quality of the synthetic microdata produced without the detailed queries would be lower than the quality of microdata that can be produced from the detailed queries, in the sense that it would be less similar to the original microdata (of course, when the goal is to produce privacy-protected synthetic microdata, being dissimilar from the original microdata in some ways is essential). So long as the complete set of statistics to be produced is known in advance and is used to produce the post-processed statistics from which the synthetic microdata are generated, this internal quality difference would never impact a published statistic.

Another drawback of the approach of using the detailed queries to produce synthetic microdata is it complicates the opportunity to release the noisy measurements. The Census Scientific Advisory Committee[4] and external experts (Dwork et al., 2021; Mervis, 2021) have all argued that releasing the noisy measurements would be valuable to users. Since the detailed queries performed for the purpose of producing synthetic microdata do not correspond to any published statistics, releasing the noisy measurements corresponding to these queries would be both confusing to users and pose an additional disclosure risk. It would not violate the formal privacy guarantee for individual microdata since these values are perturbed by privacy noise at the scale corresponding to the published privacy-loss budget (discussed more in the following section). It would, however, provide valuable detailed information to an adversary conducting an attack. This does not change the set of attacks that can be guaranteed to fail because of the formal privacy guarantee, but it could make some attacks that would be otherwise infeasible in practice now possible. The other disclosure risk with releasing the noisy measurements for these detailed queries is that they would reveal statistical information about population characteristics that are not included in any census data products. This statistical information does not disclose attributes of individuals, but may still be sensitive, such as revealing the racial composition of a particular group quarters facility.

On the other hand, not releasing the noisy measurements for the detailed queries would make the noisy measurements released for queries that do correspond to published statistics less useful and harder for users to understand. The actual published statistic cannot be derived from just the noisy measurements corresponding to published statistics. Because the post-processing adjustments done to produce those statistics also depend on the noisy measurements for the detailed queries, there would be no way to provide a transparent method for converting

[4]Census Scientific Advisory Committee (2021) exhorts, "CSAC reasserts its earlier recommendation that the Bureau release the non-post-processed data used in TDA, which are unbiased estimates. To address the Bureau's concerns that the release of such estimates would require extensive user guidance, CSAC recommends that the Bureau consider releasing such data as a research product."

noisy measurements to post-processed statistics that were created from the full noisy measurements including the detailed queries. Further, without access to the noisy measurements from the detailed queries, users who perform analyses using the available noisy measurements may end up with less accurate estimates than the ones that could be derived from the post-processed statistics, since those statistics were derived using the noisy measurements from the detailed queries.

JASON joins the Census Scientific Advisory Committee and other external experts in desiring the release of the noisy measurements, but because of the complications due to producing the synthetic microdata we instead make a weaker recommendation (**R3**) to release all noisy measurements that are used to produce a published statistic when doing so would not incur undue disclosure risk. Ideally, the Census Bureau will be able to move to a method of producing privacy-protected data products where the only noisy measurements needed are those corresponding to statistics in the published tables. Then, those noisy measurements could be released with no increase in formal privacy loss and little practical additional disclosure risk. Data users could independently post-process the data to reproduce the published tables and benefit from software tools designed to use the noisy measurements (Section 7.2). The current method of using the detailed queries to produce synthetic microdata, however, prevents this option from being possible.

5.4 Privacy Parameters

The differential privacy definition and the zero-concentrated differential privacy definition used by the Census Bureau have a privacy loss budget parameter that bounds the inference risk associated with the released data. For the differential privacy guarantees to hold for a set of collected sensitive microdata, all data released from that microdata must be released using a mechanism that satisfies the privacy definition. The overall privacy loss budget for that microdata can be determined by composing all data releases computed on that microdata. Setting a maximum privacy loss budget for a given set of microdata is a policy decision that

depends on considering the risks to individuals and to the reputation of the data curator if inferences can be made from the released data.

Because of the series of different data products to be released from the decennial microdata and uncertainty about the later data products, the Census Bureau has not yet determined a global privacy loss budget for the decennial microdata. Instead, the Census Bureau has independently determined the privacy loss budget allocated for the PL data release (United States Census Bureau, 2021d). We refer to the total privacy loss budget allocated to the PL data release as the *PL privacy loss budget*. In some Census Bureau communications this is called the *global privacy loss budget*. Calling it a global privacy loss budget is potentially misleading because it does not account for future data releases derived from the same microdata. Additional queries will need to be done on the same decennial microdata to produce the DHC data release and later data products. Since all of these data products will be publicly released and could potentially be combined by an adversary in a disclosure attack, the global privacy loss budget for the decennial microdata must cover all the data eventually released that will be derived from this microdata.

5.4.1 Setting the privacy loss budget for the PL release

The Census Bureau did a series of demonstration data releases to explore and communicate potential privacy loss budgets. Table 5-7 summarizes the privacy loss budgets considered leading up to the final production parameters used for the 2020 PL data release. The 2018 End-to-End (2018 E2E) Census Test used a PL privacy loss budget of $\epsilon = 0.25$ with pure differential privacy (Fontenot, 2019). This was divided equally among the four geographic levels (County, Tract, Block Group, Block) for which privacy noise was applied (United States Census Bureau, 2019a). The 2018 E2E test was only done for Providence, Rhode Island, there is no state-level data and county-level total population counts were treated as invariants.

To satisfy differential privacy at $\epsilon = 0.25$ requires adding a substantial amount of noise but offers a very high level of protection.

It became clear, however, that it was not possible to produce data with satisfactory utility using such a low privacy loss budget. The PL privacy loss budgets increased over the course of the demonstration data releases as summarized in Table 5-7. The Census Bureau used pure differential privacy in the early demonstration data releases and switched to zCDP (see Section 4.3.2) for the September 2020 and subsequent data releases, including the production release. The privacy loss budget was increased from $\epsilon = 4.5$ for the May 2020 demonstration data release to 12.2 for the final demonstration data release in April 2021. The final production privacy loss budget for the PL data release was set at $\rho = 2.63$. In public materials disclosing the selected privacy parameters (United States Census Bureau, 2021d), the Census Bureau does not mention the use of zCDP or the selected ρ privacy parameter, but instead publishes the privacy parameters after converting to the more familiar ϵ differential privacy notion. The zCDP privacy lost budget of $\rho = 2.63$ can be converted to (ϵ, δ)-Differential Privacy (Definition 4.3) for any choice of δ; selecting a lower δ value will result in a higher ϵ. The Census Bureau converts using $\delta = 10^{-10}$ to produce a privacy loss budget of $\epsilon = 19.61$. This means that with all but $1 - 10^{-10}$ probability, an output satisfies pure differential privacy at $\epsilon = 19.61$. With zCDP this conversion underestimates the actual privacy since there are not catastrophic failures, just a gradual increase in the risk that the pure differential privacy inference bound would be exceeded.

Other studies done in 2018–2020, including several presented to JASON for our 2019 study (JASON, 2020), considered privacy loss budgets between $\epsilon = 0.25$ and $\epsilon = 8$. The Census Bureau further evaluated their reconstruction and reidentification attack experiments on data using privacy loss budgets ranging up to $\epsilon = 16$. They found that the reidentification rate at $\epsilon = 16$ was 8.2%. This was without considering any countermeasures an adversary might use to improve results on noised data. This compared to a 17% reidentification rate when no pri-

vacy noise was applied (National Academies of Sciences, Engineering, and Medicine, 2020).

One measure for the level of privacy protection is the *Effective True Positive Rate (TPR)*. reported in Table 5-7. This measures how much the released data could benefit a worst-case adversary based on a hypothesis testing framework (Wasserman & Zhou, 2010; Kairouz et al., 2015; Balle et al., 2019; Dong et al., 2019). It considers an inference adversary who is aiming to distinguished between two possible neighboring sets of underlying microdata with equal prior probability. The metric gives a bound on the maximum true positive rate that could be achieved by an inference adversary constrained to have a maximum false positive rate, computed by this formula from Wasserman & Zhou (2010):

$$f_{\epsilon,\delta}(\alpha) = \max\{0, 1 - \delta - e^{\epsilon}\alpha, e^{-\epsilon}(1 - \delta - \alpha)\}$$

where α is the maximum false positive rate. For the values in Table 5-7, we use $\alpha = 0.01$. Without any privacy noise, the maximum TPR would be 1.0, or full confidence for the adversary. For $\alpha = 0.01$, the effective TPR corresponding to just guessing would be 0.01. As indicated by the Effective TPR values in Table 5-7, even a worst-case adversary attempting to distinguish between two possible neighboring sets of microdata would have very little advantage from receiving the produced data at the $\epsilon = 0.25$ privacy loss budget level. In other words, the best possible adversary using all available information would do at best just a tiny bit better than random guessing in distinguishing between the two source data sets differing in a way that would impact a count by just one using the released data). For the $\epsilon = 19.61$ at $\delta = 10^{-10}$ privacy parameters used in production, the effective TPR is negligibly different from 1, indicating that the formal guarantees provided by differential privacy at this level provide little guaranteed uncertainty.

The production parameters for the PL data release were set by the Census Bureau's Data Stewardship Executive Policy (DSEP) committee to provide a balance between the utility needs of stakeholders and the confidentiality requirements to

Table 5-7: Evolution of PL Privacy Loss Budget. *PL-ϵ* is the overall privacy loss budget allocated to the PL data release. The May 2020 and earlier demonstration data releases satisfy pure differential privacy (DP); later releases satisfy zero-concentrated differential privacy (zCDP). The ϵ values reported for zCDP are based on the Census Bureau's published conversions (converting the ρ parameter used at $\delta = 10^{-10}$). This is not a tight bound for zCDP, so these numbers are meaningful for interpreting the published conversions, but the actual mechanism has a lower Effective TPR. We are not aware of a precise way to estimate this for zCDP, but report the results of a search for the minimum Effective TPR found across possible ϵ, δ conversions in the bottom row. The April 2021 demonstration data release included versions using both $\epsilon = 4.5$ and $\epsilon = 12.2$, indicating that the 12.2 privacy loss budget was the one expected to approximate the final parameters. The $\epsilon = 4.5$ version was included for comparisons with previous demonstration data releases.

Date	Description	PL-ϵ	Effective TPR
Nov 2018	2018 E2E Census Test DAS (DP)	0.25	0.01285
Oct 2019	Demonstration Data Project (DP)	6.0	0.9976
May 2020	Demonstration Data (DP)	4.5	0.90018
Sep 2020	Demonstration Data (zCDP)	4.5	0.90018
Nov 2020	Demonstration Data (zCDP)	4.5	0.90018
Apr 2021	Demonstration Data (zCDP)	4.5; 12.2	0.9999955
Jun 2021	Production Parameters (zCDP; $\delta = 10^{-10}$)	19.61	0.999999997
Jun 2021	Minimum estimate for zCDP at $\rho = 2.63$	—	0.997

satisfy Title 13. An important factor in setting the production privacy parameters was the redistricting use case, and in particular the needs of the Department of Justice to enforce requirements of the 1965 Voting Rights Act. Section 2 of the Voting Rights Act prohibits voting practices that discriminate on the basis of race, which includes designing voting districts to unfairly decrease the political power of minority groups (United States Code, 1965). An example of a case the Department of Justice brought under the Voting Rights Act is *United States v. State of Texas* 2013 (State of Texas, 2013). In this case the Department of Justice argued that the redistricting plan generated by the Texas Senate and House Redistricting Committees split precincts more than necessary to equalize the sizes of voting districts, but in a way that diluted minority voting strength and deliberately avoided creating

any districts where the majority of the voting age population in the district were minorities. Such cases depend on accurate information on race, down to low-levels of geography, as provided by the PL data release. We provide more detail on the experiments used to test the impact of the Top-Down Algorithm on these statistics in Section 6.2.

JASON's intent in pointing out the ranges of privacy loss budgets considered up until the final production parameters were set is to recognize the Census Bureau's openness of its process in determining the disclosure avoidance mechanisms. The increases in the privacy loss budget were a natural result of making changes in response to stakeholder's feedback on utility problems. The privacy loss budgets used in the experiments covered a reasonable range and the choices used in the demonstration data products were well justified. It is to the Census Bureau's credit that these data were provided to allow stakeholders to analyze it and critique its fitness for their uses.

5.4.2 Allocating the Privacy Loss Budget

In addition to allocating the overall privacy loss budget to the PL data release, parameters of the Top-Down Algorithm determine how that privacy loss budget is partitioned across all of the queries used to produce the synthetic microdata from which the PL data release are produced. The first split is between the persons file and the housing units file. For the PL queries, these can be considered separately. For the PL data release, there is only one query that uses the housing units file to produce a count of the number of occupied housing units in each geographic unit (the H1 table). Since this depends on the same underlying microdata, however, parallel composition does not apply and privacy loss budget must be allocated to this query. For the production parameters, 2.6% ($\rho = 0.07$) of the privacy loss budget is allocated to this leaving $\rho = 2.56$ for the rest of the queries. This is all done on the persons file to produce the P1–P5 tables in the PL data release (see Table 2-1).

The Top-Down Algorithm allows the privacy-utility tradeoffs to be controlled at the level of individual queries by how the overall privacy loss budget is partitioned. A simple strategy would just divide the privacy loss budget equally across the geographic levels and then among the different queries at each level. This approach was used in the early demonstration data releases. Based on the Census Bureau's experiments and feedback from stakeholders, this evolved to the imbalanced privacy loss distribution shown in Figure 5-3. That distribution was set primarily based on the utility requirements for the redistricting use cases (see Section 6.2).

Figure 5-3 gives the percentage of the $\rho = 2.56$ privacy loss budget allocated to the persons file for the PL data release. For example, the value 9.628 in the cell corresponding to the *Detailed* query level means that 9.628% of the privacy loss budget ($\rho \sim 0.246$) is allocated to that query. The value 21.443 in the *All* level for the *Detailed* query means that across all the geographic levels the detailed queries are using over 21% of the privacy loss budget is allocated to the persons file. The noisy measurements resulting from these queries are used to improve the accuracy of published statistics and to aid the microdata generation, but no published statistics correspond directly to these queries.

The large shares of the privacy loss budget allocated to the Hispanic $\times$ Race queries at the Tract and Optimized Block Group levels reflect the importance of these values for enforcing the Voting Rights Act.

	US	State	County	Tract	OBG	Block	All
Total Population	0.000	32.352	8.321	6.403	12.746	0.005	59.826
Race	0.032	0.051	0.027	0.033	0.022	0.009	0.174
Hispanic	0.016	0.051	0.027	0.020	0.022	0.005	0.142
Voting Age	0.016	0.051	0.027	0.020	0.022	0.005	0.142
HHInstLevels	0.016	0.051	0.027	0.020	0.022	0.005	0.142
Household/GQ	0.016	0.051	0.027	0.020	0.022	0.005	0.142
Hispanic × Race	0.081	0.103	0.075	7.898	7.887	0.021	16.063
Voting Age × Race	0.081	0.103	0.075	0.082	0.067	0.021	0.428
Voting Age × Hispanic	0.016	0.051	0.027	0.020	0.022	0.005	0.142
Voting Age × Hispanic × Race	0.274	0.292	0.269	0.274	0.179	0.070	1.357
Detailed	1.990	1.972	2.007	1.969	9.628	3.876	21.443
Total	2.537	35.131	10.905	16.760	30.642	4.025	100.000

Figure 5-3: Allocation of PL Privacy Loss Budget. Each cell is the percentage of the persons file privacy loss budget allocated to the given query. *OBG* is an abbreviation for *Optimized Block Group*. The *Detailed* queries are the Household or Group Quarters Type (8) × Voting Age (2) × Hispanic/Latino origin (2) × Race (63) queries that are used to produce the synthetic microdata. The *All* column is the share given to that query across all geographic levels, and the *Total* row is the total share allocated to that geographic level. The value of 32.352 in the State total population cell reflects privacy noise used for the total populations for all AIAN areas within the state for the 36 states that include tribal areas. Values are derived from United States Census Bureau (2021g).

This Page Intentionally Left Blank

6 EVALUATING THE DAS

In this section we evaluate the impact of the disclosure avoidance system on the utility of the released data products. First, we consider the uncertainty in census data apart from the intentional noise introduced for disclosure avoidance. Section 6.2 summarizes experiments to measure the impact of privacy noise on accuracy. Section 6.3 describes experiments JASON did to understand the impact of post-processing on accuracy and bias.

6.1 Accuracy and Uncertainty

The concept of accuracy is at the heart of users' concerns over the impact of noise addition for privacy in data products released by the Census Bureau. Several stakeholder groups have expressed concerns that Census methods for protecting privacy could result in amplifying undercounts and shifts in population. Consequences of inaccuracies in Census data could include loss of political representation and misappropriation of federal benefits received by some groups or the inability of local jurisdictions to properly allocate goods and services within a community.

The definition of accuracy takes on different meanings depending on the context. The International Statistical Institute (2003) defines *accuracy* as the "closeness of computations or estimates to the exact or true values that the statistics were intended to measure". In the context of the decennial census enumerations before any disclosure avoidance adjustments, accuracy refers to the closeness of the reported enumerations to the true values. By contrast, in the context of the discussion around data releases modified to ensure differential privacy and subsequent post-processing to satisfy constraints, accuracy often means the difference between the reported statistics as compared to those same statistics as computed from the CEF without any privacy noise. One way to evaluate the impact of disclosure avoidance mechanisms would be to compare the variance introduced

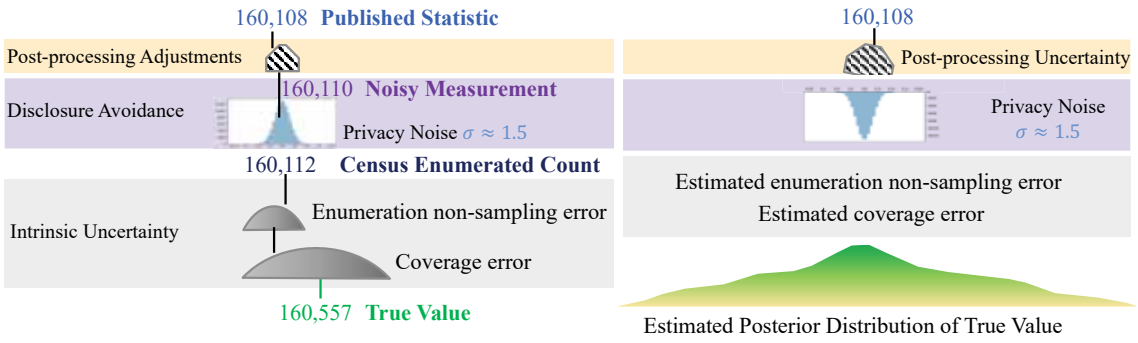

Figure 6-1: Accuracy from the perspective of the data producer (Census Bureau) and data consumer. (The 160,112 census enumerated count is the population of Chattanooga City, Tennessee as enumerated by the 2010 census (Table 6-9), but is just used as an arbitrary example. All the other values and scales are for illustrative purposes only, and do not represent real data.)

by the mechanisms to the uncertainty inherent in a census enumeration due to the enumeration process itself.

Figure 6-1 illustrates how accuracy appears from the viewpoint of the Census Bureau as the data producer (left side), and from the viewpoint of a data user (right side). The Census Bureau starts the data release process from the enumerated count, which is its estimate of the true value. The actual true value of most statistics released by the Census Bureau is unknown and the enumerated count reflects uncertainties in the data collection process itself. If privacy were not a concern, this value could be released directly. To avoid sensitive disclosure and satisfy the desired formal privacy properties, randomly sampled noise from a publicly known noise distribution is added to this value. As discussed in Section 5, this produces the *noisy measurement*. Following this, further post-processing adjustments are done to produce the final published statistic. From the data producer's view, *accuracy* is most directly measured as the difference between the published statistic and the enumerated count since both values are known. This measure of accuracy, however, does not account for the intrinsic uncertainty in the enumerated count.

From a data consumer's viewpoint, the published statistic is the starting point. Many users of census data products view this number as the one true number, and do not consider its uncertainty. More sophisticated users accept that this published statistic is an estimate that includes various sources of uncertainty, and will want to estimate a posterior distribution of the true value from the published statistic and available information about the uncertainty in that estimate. Privacy noise is just one of the causes of uncertainty and it is the one that can most easily be accounted for since the probability distribution used to produce the privacy noise is known and has well understood mathematical properties, including being balanced. The other sources of uncertainty, including post-processing and the error in the enumerated count, are less clear to a data consumer and may introduce systematic biases.

To provide more background for differences in these interpretations of accuracy, we examine how these are quantified through demographic analysis, post-enumeration surveys, and a simulation exercise designed to capture the intrinsic uncertainty.

6.1.1 Demographic analysis

Since 1960, the Census Bureau has applied demographic analysis (DA) to estimate the national population on Census Day, April first of the first year in a new decade. A range of estimates are released in December of the year the census is taken, several months before release of the official census enumeration for the national population and apportionment of congressional districts.

Records of births, deaths, and Medicare enrollments are used to estimate population changes since the previous census. Because immigration records are incomplete, the annual American Community Surveys (ACS) are used to estimate net changes in international residents. Age and sex are included, but racial information is limited. For example, earlier birth records specified only white and

non-white. Additional racial categories used in recent censuses allow for more detailed estimates for young people in the 2020 DA release.

Demographic analysis results in three official sets of estimates for each statistic which represent low, middle, and high estimates based on different data sources and methods of estimation. The estimates for the 2010 national population were 305.648 million (low), 308.475 million (middle), and 312.713 million (high); the corresponding enumeration was 308,745,538. Demographic estimates of the 2020 national population were 330.730 million (low), 332.601 million (middle), and 335.514 million (high), compared to 331,449,281 for the official enumeration (Jensen et al., 2020). The middle estimate was much farther from the enumeration than in 2010, as were related demographic projections of state populations. This can cause political controversy when changes in state apportionments using the enumerations differed substantially from expectations based on demographic analysis (Wall Street Journal Editorial Board, 2021).

In addition to demonstrating that the Census Bureau stands behind its data, demographic analysis is used to identify potential correlation biases in the enumerations, such as groups missed in the census enumeration and in the post enumeration survey. Revised DA estimates are released the year following the census. For example, using improved processing and data, the middle estimate was later revised to 308.346 million.

6.1.2 Post-enumeration surveys

Post-enumeration surveys have been used to assess the coverage of the census enumeration. This includes potential overcounts and undercounts in the actual enumerations calculated from the CEF. Because procedures vary over the decades, post-enumeration surveys have had different names for different decennial censuses. The one for the 2010 census was called the Census Coverage Measurements (CCM) Program (Mule, 2012a,b).

In parallel with preparations for the 2010 census and guided by earlier censuses, in October 2008 the CCM program selected 12,500 sample block clusters, a subset of those used in the census enumeration. This process was designed to be completely independent of the actual decennial enumeration and to include locations that are difficult to enumerate, such as states with small population. Next, a list of addresses of housing units in sampled block clusters was created, again independent of those chosen for enumeration. Determination of who lived in the housing units on Census Day (1 April 2010) followed soon after census enumeration. The CCM provided an independent population sample (*P-sample*) for statistical comparison with the official census enumerations in the block groups used by CCM (*E-sample*). Two estimates of the same block clusters allowed the application of Dual System Estimation, a method based on logistic regression, to estimate four components of census coverage: correct enumerations, erroneous enumerations, whole-person census imputation counts, and omissions. Estimation was done for the fifty states, the District of Columbia, and Puerto Rico. Group housing (dormitories, barracks, nursing homes, and jails) was excluded from the analysis, as well as remote regions in Alaska.

Table 6-8 summarizes the 2010 component estimates. The overall coverage measure is net undercount, given by

$$\textit{Net Undercount} \equiv \textit{DSE} - \textit{Census},$$

where DSE is the population count estimate produced by the Dual Systems Estimation. In 2010, for the national population, the result was a statistically insignificant net overcount of 0.01% (hence a negative net undercount) with a standard error of 0.14%. As a measure of uncertainty, however, this is misleading, as it results from near-cancellation of much larger uncertainties, indicated by 3.3% of erroneous enumerations.

Table 6-9 is an example of estimates of total population within a single state, Tennessee. Thirteen counties containing 60% of Tennessee's total population were

Table 6-8: Comparison of 2010 census counts and CCM estimates (in thousands), from Mule (2012b). To reflect the CCM procedure, this comparison does not include persons in group quarters or in remote Alaska enumeration regions (hence, the census count is 300.7 M, not the total population count of 308.7 M). Omissions are persons who should have been counted but were not. Many of these may have been included in census imputations.

Coverage component	**Estimate**		**Standard Error**	
	Persons	Percent	Persons	Percent
Census Count	300,703	100.0	-	-
Correct enumerations	284,668	94.7	199	0.07
Correct block cluster	280,852	93.4	220	0.07
Correct county, wrong block cluster	2,039	0.7	55	0.02
Correct state, wrong county	830	0.3	34	0.01
Wrong state	948	0.3	31	0.01
Erroneous enumerations	10,042	3.3	199	0.07
Duplications	8,521	2.8	194	0.06
Other reasons	1,520	0.5	45	0.01
Whole-person imputations	5,993	2.0	0	0
Census Coverage Measurements	300,667	100.0	429	
Correct enumerations	284,668	94.7	199	0.1
Omissions	15,999	5.3	440	0.1
Net Undercount	−36	−0.01	429	0.14

sampled during the CCM, along with six cities, each in one of the twelve counties sampled. Net undercount for the state was estimated at 0.12%, ten times that for the country. County undercounts varied from −0.63% to +1.00%. All city undercounts were positive, ranging from +0.40% to +1.42%. Accuracy coverage measures that take into account the sampling uncertainty are reflected in the 90% confidence intervals for the estimated undercounts. These are often larger by a factor of 10 or more as compared to the national aggregate figures. The implication is that some cities and counties may be receiving federal benefits that are several percent higher or lower than they would be if their true population were determined and used instead of the census estimate.

Table 6-9: Census Coverage Measurement (CCM) estimates of net undercounts and the equivalent percentage of the DSE for selected counties and cities in Tennessee for the 2010 Census. The upper part of the table lists cities and their encompassing counties that were surveyed. Nashville–Davidson is the part of Nashville City that is in Davidson County. The lower part lists smaller counties that were surveyed.

Unit	Population	Undercount	90% Confidence
State of Tennessee	6,192,633	+0.12%	−1.78% to 2.01%
Chattanoga City	160,112	+0.59%	−4.17% to 5.35%
Hamilton County	326,685	+0.04%	−4.03% to 4.11%
Clarksville City	130,008	+1.10%	−3.96% to 6.16%
Montgomery County	168,915	+0.89%	−3.88% to 5.66%
Knoxville City	168,826	+0.40%	−4.32% to 5.12%
Knox County	419,878	−0.05%	−3.91% to 3.80%
Memphis City	630,353	+1.42%	−2.22% to 5.05%
Shelby County	909,315	+1.00%	−2.29% to 4.30%
Murfeesboro City	104,321	+0.44%	−4.79% to 5.68%
Rutherford County	257,495	+0.09%	−4.19% to 4.38%
Nashville–Davidson	575,429	+0.82%	−2.75% to 4.39%
Davidson County	600,811	+0.77%	−2.76% to 4.31%
Blount County	120,983	−0.36%	−5.46% to 4.75%
Sullivan County	154,192	−0.63%	−5.49% to 4.23%
Sumner County	159,393	−0.32%	−5.11% to 4.48%
Washington County	118,330	−0.35%	−5.49% to 4.79%
Williamson County	182,029	−0.23%	−4.86% to 4.41%
Wilson County	112,761	−0.27%	−5.45% to 4.90%

Figure 6-2 shows the 90% confidence intervals for the net undercounts for the states. State undercounts were 10 to 100 times larger than the national undercount of 0.01% and mostly bounded by ±1%. Overcounts (that is, negative undercounts) were common, with 31 states having overcounts. However, zero is included in the 90% confidence interval for every state.

Here we have focused on CCM estimates of population coverage. Over the nation, state over and undercounts of the states largely cancel out. CCM also estimated the coverage of housing units and obtained similar uncertainties. The correct census enumerations for the housing units were 97.3%, with 96.1% in the correct block cluster, 1.2% in nearby blocks, and 0.1% somewhere else (Mule, 2012a). The CCM estimated correct housing unit enumerations as 96.8% and omissions as 3.2%.

The patterns of coverage mismatches have helped the Census Bureau make improvements to their enumeration processes. The post-enumeration survey processes have also evolved over the decades. Between 2000 and 2010, the decision to measure the components shown in Table 6-8, versus just measuring undercounts or overcounts was developed to better capture what was contributing to the differences (National Research Council, 2009).

The post-enumeration surveys are designed to estimate the accuracy in the Census as enumerated. The *E-sample* in the comparison is based on the CEF, without any disclosure limitation adjustments. The Census Bureau could take this measurement a step further to address accuracy in the context of data releases modified to ensure differential privacy and subsequent post-processing. Recognizing the coverage estimates based on comparisons to the CEF are most useful for the Census Bureau internally in evaluating the methods it uses to enumerate the population, it would help users understand the potential net undercounts in the data they actually use if similar estimates were done in comparison to the publicly released data which includes the privacy noise and post-processing adjustments.

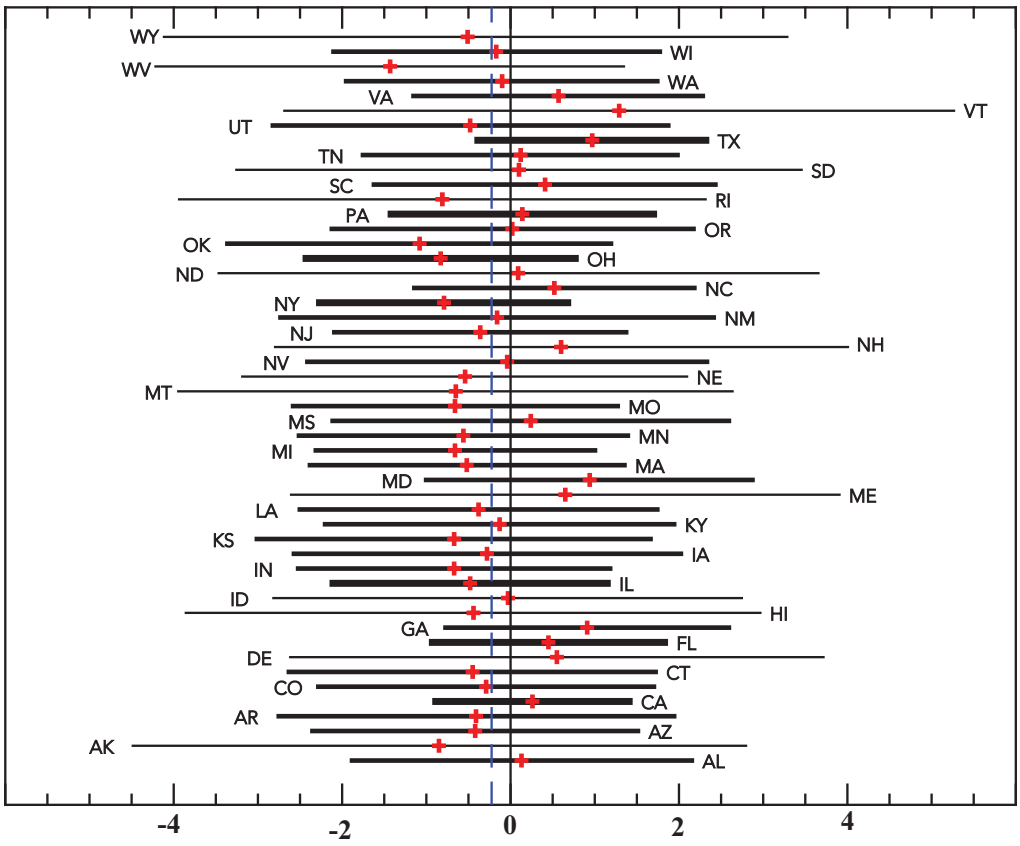

Figure 6-2: Undercounts for the fifty states from the 2010 Census Coverage Measurements (CCM). The letters are the state abbreviations. The horizontal lines are 90% confidence intervals for the undercount. The thickness of the lines corresponds to the total state populations, with thinnest (populations < 2 M), medium (populations between 2 M and 10 M) and thickest (populations over 10 M). The plus signs are the median (net undercounts) undercounts for each state. The blue dashed horizontal line is the median undercount (−0.23) across all 50 states.

6.1.3 Intrinsic variability in the enumeration process

The demographic analyses and post-enumeration surveys do not account for intrinsic variability that could exist in the actual enumeration produced in the CEF. Whereas the population counts from a decennial census do not contain sampling error, they do include non-sampling sources of errors, such as those identified through CCM (Table 6-8). This means that if the Census were repeated on the exact same population using the same methods, the results would vary. The question is, by how much?

Estimating the magnitude of this uncertainty would provide a sense of how variable the data resulting from the enumerations are before any addition of privacy noise. Intuitively, the impact of the privacy noise depends on the magnitude of the uncertainty in the underlying data. If the data already have intrinsic variability, it could be that the impact of the privacy noise on overall uncertainty is negligible. If the underlying data are nearly exact matches for the true values, the only significant uncertainty would be due to the privacy noise.

The Census Bureau recently developed a simulation framework to estimate this intrinsic variability and applied it to the 2010 Census (Schafer, 2021). Block-level simulations were run based on CCM coverage patterns to simulate the variability in the actual enumeration by mimicking the repeating of the census, while holding the population fixed in the same housing units.

For the simulation run-to-run, a Bayesian Poisson model was used to inject substitutions into a block using a Dirichlet prior distribution with shape parameters based on the CCM estimates. Each simulation used the same percentage of correct enumerations, on the conservative assumptions that these persons would have been as easy to enumerate during the simulated runs as during the actual census.

Results, based on 100 simulations, are expressed as mean absolute percent error,

$$MAPE \equiv 100 \times \left(\frac{|N_D^* - N_D|}{N_D} \right) \tag{6-1}$$

where N_D is the published 2010 census count in domain D and N_D^* is the simulation estimate. For the national estimate, the *MAPE* is 0.002%. The smallest state estimate was 0.006% for New York, only three times larger than the national estimate, but the largest state estimate was 0.077% for Maine. Counties exhibit a similar increase in *MAPE* with decreasing size. The average for all counties, 0.31%, was bounded by 0.08% for counties with populations of one million or greater and 1.60% for counties with fewer than 1,000 persons. The county *MAPE* estimates are similar to the net undercounts of the Tennessee counties sampled by the 2010 CCM (Table 6-9).

A natural next step would be to compare the magnitude of these uncertainties to the additional uncertainty resulting from the privacy noise plus post-processing adjustments that produce the published counts. This would enable a deeper understanding of which reported statistics are overly influenced by the disclosure avoidance process and which have substantial uncertainty only because of disclosure avoidance.

6.2 Utility Experiments

The Census Bureau conducted experiments to measure the impact of the DAS on accuracy. For each demonstration data release, a variety of accuracy metrics were included. Hawes & Spence (2021) provide a summary of accuracy metrics for demonstration releases leading up to the final production parameters. Key metrics reported include the mean absolute error (MAE) and mean absolute percent error (MAPE), across a variety of queries. These metrics show significant improvements for nearly all queries between the early demonstration data releases and the final production parameters. The accuracy (relative to the non-noised statistics)

improvements are due to increases in the privacy loss budget and improvements to the Top-Down Algorithm. Other changes, such as the optimized spine and decisions about how to allocate the privacy loss budget across queries and geographic levels improve some results, but may increase the error for others. As an example of the amount of improvement in accuracy in these analyses, the mean absolute error for the total population in a county decreased from 76.5 persons for the October 2019 demonstration data (pure differential privacy with $\epsilon = 6$) to 1.3 persons for the production settings.

Average errors are a useful measure for revealing the overall impact of the DAS, but to evaluate fitness for particular use cases requires more detailed analysis of the distribution of those errors. Because of its importance to Voting Rights Act cases, Wright & Irimata (2021a) studied the impact of the Top-Down Algorithm on the statistics for the population of the largest demographic group (race and Hispanic or not ethnicity) in a geographic unit. The original study investigated the impact of the privacy parameters used for the April 2021 demonstration data, where a privacy loss budget of $\epsilon = 10.3$ was allocated to the PL queries on the persons file (out of $\epsilon = 12.2$ total, the remainder being used for the housing units file). The main criteria for their analysis was the difference of the ratios between the value of a statistic produced using the Top-Down Algorithm with that value produced using the swapping methods used for the 2010 census.[5]

For a given demographic group, such as a minority race, the differences in ratios was defined as

$$DR_{group} = \left| \frac{C_{SWA}(group)}{C_{SWA}(all)} - \frac{C_{TDA}(group)}{C_{TDA}(all)} \right|,$$

where C_{SWA} is the value of the statistic produced when swapping is used and C_{TDA} is the value of a statistic when the Top-Down Algorithm is used.

[5]The decision to compare to the swapped counts rather than the raw CEF micro data was done to allow public reproducibility and avoid the need to consider disclosure risk on the results of the study. The swapping was done in a way that preserved total and voting age population counts at all geographies (but not necessarily by-race counts).

This metric captures what is most important for the Voting Rights Act, which is the population of the largest group within a geographic unit. The main test for fitness for use for this purpose is to determine for what fraction of the geographic units considered is the value of $DR_{largest} \leq T$, where *largest* is the group in the geographic unit with the largest population (as given by the statistics using the swapping method), and T is some accuracy threshold such as 0.05.

Wright & Irimata (2021a) looked at the number of block groups for which $DR_{largest}$ exceeded 0.01, 0.03, and 0.05, categorized by the size of the total ($C_{SWA(all)}$) population of the block group. They found that the Top-Down Algorithm results were unreliable for small block groups, with more than half of the block groups of size less than 200 having $DR_{largest} > 0.05$. Such small block groups account for only 351 of the total 199,698 block groups across the United States included in their study. For over 95% of block groups with population at least 550, $DR_{largest} \leq 0.05$. This same result was confirmed over 25 runs of the Top-Down Algorithm, using different randomness. This provides high confidence that any run of the Top-Down Algorithm will produce reliable counts for the largest demographic group in areas with size over 550.

The original experiments in Wright & Irimata (2021a) used the version of the Top-Down Algorithm and parameters for the April 2021 demonstration data release ($\epsilon = 12.2$ of which 10.3 was allocated to the persons file, which is the source of all the data used). Wright & Irimata (2021b) updated the experiments using the final production parameters (which increase the privacy loss budget allocated to the persons file to $\rho = 2.56$ (reported as $\epsilon = 17.14$) and also adjusted the per-query allocations). As summarized in Table 6-10, these changes improved the accuracy of the relevant statistics, supporting a conclusion that the statistics published in the PL data release are reliable for redistricting purposes for over 95% of block groups with total populations above 450. This includes over 98.5% of block groups and an even higher fraction of the population.

Table 6-10: Summary of data from Wright & Irimata (2021a) (the 2021-04-28 columns, corresponding to the April 2021 demonstration data release) and Wright & Irimata (2021b) (*Production* columns, using the final Top-Down Algorithm settings). This table combines data from Table 3 in each of the two reports, showing improvement in reliability with the production Top-Down Algorithm. The Count is the number of block groups in the size range and the next column gives the cumulative fraction of blocks covered by groups up to that size (out of 199,698 total block groups). The highlighted rows indicated where the 95% reliability threshold is crossed for the Production (magenta) and April (cyan) data.

C_{SWA} Range	Count	Fraction Covered	$DR_{largest} \leq 0.01$		$DR_{largest} \leq 0.05$	
			2021-04-28	*Production*	2012-04-28	*Production*
50–99	128	0.0006	0.117	0.125	0.406	0.516
100–149	99	0.0011	0.091	0.182	0.465	0.707
150–199	124	0.0018	0.113	0.169	0.557	0.758
200–249	154	0.0025	0.214	0.266	0.714	0.792
250–299	209	0.0036	0.211	0.292	0.713	0.857
300–349	264	0.0049	0.212	0.364	0.780	0.890
350–399	407	0.0069	0.233	0.337	0.843	0.870
400–449	569	0.0098	0.290	0.408	0.896	0.932
450–499	915	0.0144	0.327	0.409	0.936	**0.955**
500–549	1699	0.0229	0.343	0.420	0.937	0.959
550–599	3238	0.0391	0.381	0.455	**0.958**	0.965
600–649	5131	0.0648	0.396	0.458	0.972	0.975
650–699	6683	0.0982	0.420	0.472	0.975	0.975
700–749	7356	0.1351	0.447	0.501	0.983	0.983
750–799	8170	0.1760	0.448	0.516	0.984	0.985
800–849	8213	0.2171	0.479	0.527	0.991	0.990
850–899	8441	0.2594	0.497	0.552	0.989	0.991
900–949	8657	0.3027	0.502	0.556	0.993	0.992
950–999	8723	0.3464	0.520	0.585	0.995	0.995

Wright & Irimata (2021a) conducted their experiments using the full Top-Down Algorithm, including both the privacy noise and the post-processing steps, and were not able to separate the impact of the privacy noise needed to satisfy the formal privacy requirements from the post-processing used to ensure non-negativity and consistency as needed enable generation of microdata.

Wright & Irimata (2021a) speculated that post-processing resulted in more uncertainty than the privacy noise, but were not able to quantify the impact of each separately:

> *"The variability and uncertainty due to the activity of the second component [post-processing] is less well understood by us, and we believe it currently contributes more variability and uncertainty than the first component. We believe that the empirical variability reported in this study is an overall combination of variability and uncertainty from the two components."*

6.3 JASON's Experiments

The questions raised about the separate impacts of privacy noise and post-processing, including by Wright & Irimata (2021a), motivated JASON to conduct experiments to understand and quantify the additional impact of the post-processing on accuracy and biases in the results. Before we explain how the experiments were done and present our results, there are two important caveats:

1. We did not have access to the Census Bureau's Top-Down Algorithm implementation or attempt to fully replicate all the complexities in the Top-Down Algorithm. Instead, we developed a much simpler algorithm that aims to capture the most important aspects of the Top-Down Algorithm—adding privacy noise and enforcing aggregate constraints and non-negativity constraints.

2. We did not conduct our experiments on the full census queries or data, but only on a single query (total population by county) and only for a subset of counties.

Hence, our results do not provide a direct comparison to the results in Wright & Irimata (2021a) or other Census Bureau reports, and the actual errors and biases we find may be higher than what would be expected with the full Top-Down Algorithm which includes several additional measures designed to reduce these biases. Nevertheless, we believe that our algorithms are close enough to the ones used in the DAS that the results of our experiments are useful for understanding important trends and tradeoffs, and hope that the design and interpretation of our experiments will be useful to the Census Bureau in considering experiments that could be done using the full implementation of the Top-Down Algorithm and data available to the Census Bureau.

6.3.1 Solving constraints

This section introduces the simplified version of the Top-Down Algorithm JASON implemented for our experiments. It captures essential aspects of the algorithm, but does not attempt to reproduce all the complexities in the algorithm used by the Census Bureau. Instead, our algorithm, which we call the *Reasonably Good Algorithm* (RGA), implements the privacy noise of the Top-Down Algorithm,[6] and then adjusts the results of the noisy measurements to satisfy two kinds of constraints: *aggregate constraints*, enforcing that sums over subsets of the statistics must be equal to the statistic that corresponds to their sum (e.g., sum of the populations of all counties in a state must equal the state population), and *non-negativity constraints*, requiring that none of the reported counts can be negative (which can be viewed as an inequality constraint requiring that a value is ≥ 0).

In the Top-Down Algorithm, these constraints are satisfied using a least-

[6]We add discrete Gaussian noise using the implementation provided by Canonne et al. (2020) from https://github.com/IBM/discrete-gaussian-differential-privacy.

squares optimizer capable of satisfying equality and inequality constraints. The optimizer looks for a solution that minimizes the L_2 norm relative to the noised data while satisfying the various constraints. In the experiments below we also use an optimizer to solve the constraints. The particular optimizer is the SciPy function scipy.optimize.minimize (Jones et al., 2001). It is invoked using the Sequential Least Squares Programming (SLSQP) method. Note that unlike the Top-Down Algorithm which uses controlled rounding to produce integral outputs, this optimizer does not yield integer results and the RGA does not round the outputs to integers. We present an alternative way of solving the constraints analytically in Section 6.3.4.

6.3.2 Bias from post-processing

We conducted experiments with the RGA to understand the consequences of the aggregate and non-negativity constraints in the post-processed results. For our experiments, we use a single query on data released from the 2010 Census: the voting age Hispanic population by county. This query reports a total of 58,479,383 people distributed across 3,142 counties, encompassing all 50 states.

The voting age Hispanic population for all 3,142 counties is shown in Figure 6-3, as a function of the rank of the population. Figure 6-4 shows an example of the noise sampling at $\sigma = 200$, which is much more noise than is used in typical queries but is useful for visualization. The effect of this random noise on our trial query, voting age Hispanic population by county, is shown in Figure 6-5. The black line is the original distribution from Figure 6-3. The blue dots indicate noised population values when the noised population is non-negative; the red dots indicate values when the noised population is less than zero. Note that the red dots only occur for small, low-rank, populations. It is these red dots which the non-negativity constraint addresses.

The upper panel of Figure 6-6 shows the difference between the noisy mea-

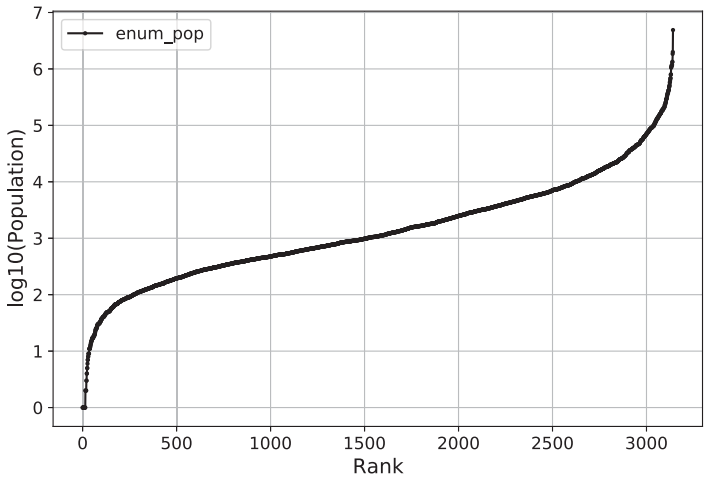

Figure 6-3: Population of Hispanic voters in the 3,142 counties ($\log_{10}$), as a function of the rank of the population.

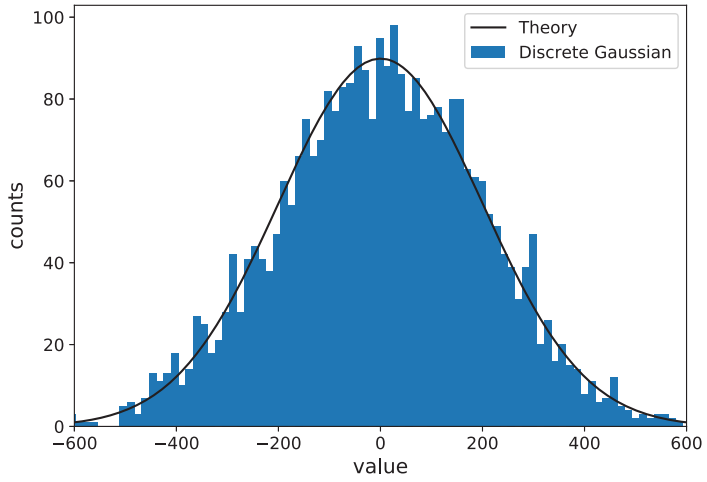

Figure 6-4: Histogram of the sampled discrete Gaussian noise added to the population query. The standard deviation of the parent distribution is 200. The solid line represents the theoretical continuous Gaussian distribution.

surement (pop_n) and the enumerated population (pop_enum) for each county. The abscissa is in units of σ, which is 200 in this example, selected for visualization not because it is a typical noise scale. The bottom panel shows the post-processed results after the both the aggregate and non-negativity constraints are satisfied (pop_nip). The green line shows the negation of the population of the county (normalized by σ), which is the lowest value the noise can be after adjusting to satisfy the non-negativity constraint.

This figure illustrates the impact of the post-processing adjustments. For counties with populations less than a few σ, the non-negativity constraint results in a positive bias towards values that are greater than the enumerated population. Because of the aggregate constraints, all of the count adjustment due to this must be removed from the larger counts, so those counties exhibit a negative bias towards values lower than the enumerated population. The amount of the bias depends on the scale of the noise relative to the distribution of the values for the query.

6.3.3 Measuring the bias

In this section, we look deeper into the bias. We solve the constraints for each state, individually. The output statistics are the Hispanic voting-age population by county for each state, with an aggregate constraint that the sum of all the counties in a state must equal the total count for the state and a non-negativity constraint for each count. We solve the problem 100 times for each state, each time using a different sampling of the noise. Thus, for each county, we have a distribution of populations for which we calculate a measure of the distribution on the biases.

We calculate the squared error of the post-processed county population $\widehat{p}$ relative to the enumerated population in the county p,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p}_i - p)^2 \tag{6-2}$$

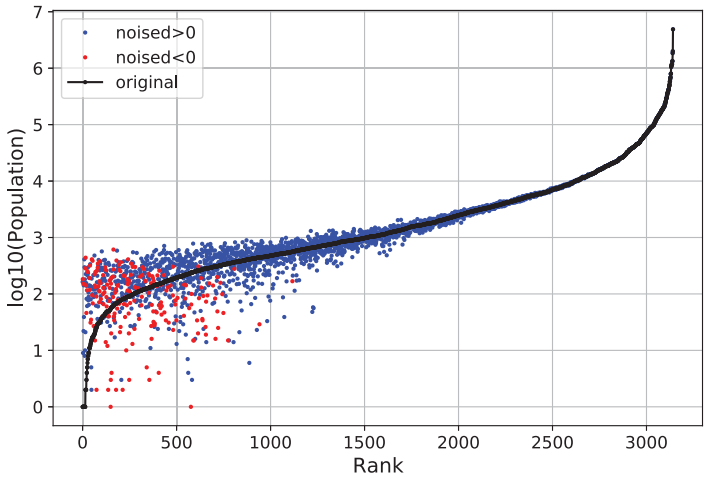

Figure 6-5: The effect of the discrete Gaussian noise on the county populations. The black line is the original distribution. The blue dots indicate values where the noisy measurement is non-negative. The red dots indicate the negation of the noisy measurement when the value is negative. Note that the red dots only occur for small populations.

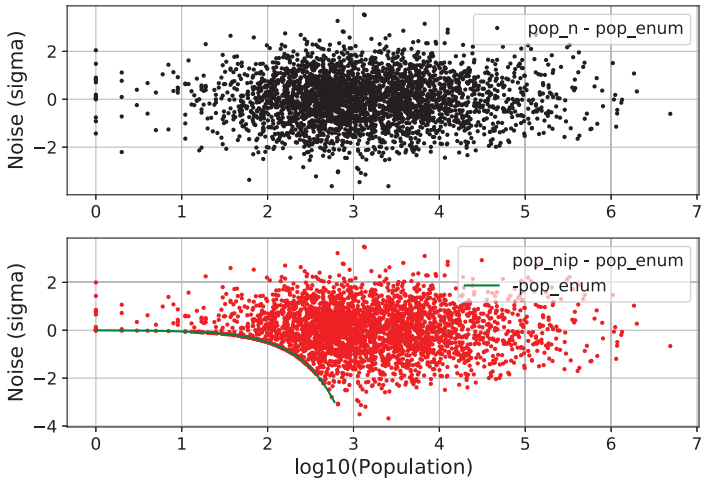

Figure 6-6: The upper panel shows the difference between the noisy measurement for each county population and its enumerated value; bottom panel the post-processed results. The horizontal axes are in units of σ (= 300). The green line is the negation of the population of the county (normalized by σ). This defines the most negative the noise can be after satisfying the non-negativity constraint.

where the sum is over the N executions of the algorithm with different samplings of the noise. For our experiments, we use $N = 100$.

We can separate this into two separate terms that split the positive and negative adjustments:

$$\sigma_+^2 = \frac{1}{N} \sum_{\hat{p}_i \geq p} (\hat{p}_i - p)^2, \qquad \sigma_-^2 = \frac{1}{N} \sum_{\hat{p}_i < p} (\hat{p}_i - p)^2$$

Clearly, $\sigma^2 = \sigma_+^2 + \sigma_-^2$. For symmetric noise, such as a discrete Gaussian distribution, without any post-processing adjustments it should be the case that $\sigma_+^2 \approx \sigma_-^2$.

We can quantify deviation away from this symmetry by the expression:

$$B = \frac{\sigma_+^2 - \sigma_-^2}{\sigma_+^2 + \sigma_-^2} \tag{6-3}$$

This measure of bias, B, can take values in the range $[-1, 1]$. If the noise distribution is symmetric, then $B \approx 0$; if all adjustments are positive, $B = 1$.

Figure 6-7 shows the calculated bias B for every county, as a function of the population of the county. The vertical dashed blue line indicates 200, which is the σ of the noise added to the raw query. The solid red line is an average calculated over a window that includes the nearest 100 points, as measured by population. Note that all populations below σ are biased towards positive values. Correspondingly, the invariant condition then enforces that the populations larger than σ have a small negative bias. The non-negativity constraint tilts distributions such that small populations have a slight positive bias, while larger distributions have a slight negative bias. Figure 6-8 shows the same calculation, but using $\sigma = 38$, which is more representative of typical values used in the Top-Down Algorithm. The blue line now indicates a county population of 38. The solid red line is an average calculated over a window that includes the nearest 100 points, as measured by population. Once again, we see there is bias for counties with populations less than σ.

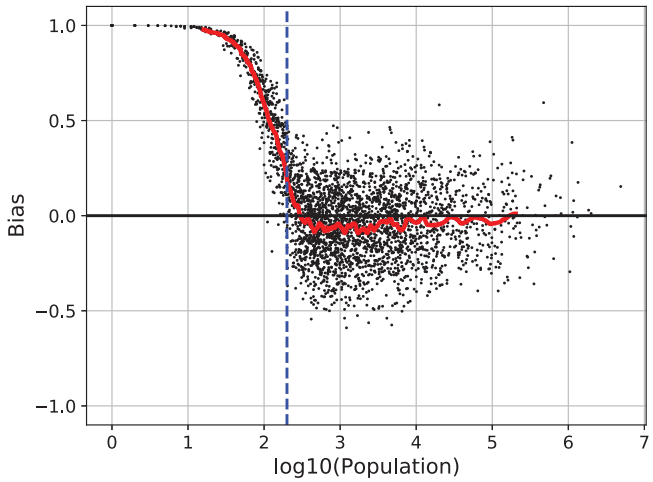

Figure 6-7: The calculated bias, $B = (\sigma_+^2 - \sigma_-^2)/(\sigma_+^2 + \sigma_-^2)$, calculated for every county, as a function of its Hispanic voting age population. The vertical dashed blue line indicates the value of σ for the noise distribution, which is $\sigma = 200$ for this graph. The solid red line is an average calculated over a window that of the nearest 100 points by population. Note that for all populations less than σ are biased towards positive values, resulting in larger post-processed values than the enumerated ones. Correspondingly, the non-negativity constraint combined with the aggregate constraint means that the populations larger than σ have a small negative bias.

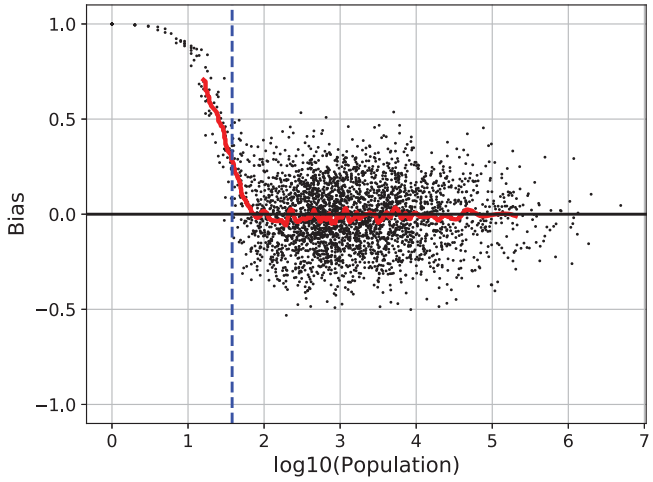

Figure 6-8: Same graph as Figure 6-7, but with $\sigma = 38$.

To visualize how these biases might impact particular counties, we simulated results for two counties in Virginia: Fairfax County (with 187,160 Hispanic voting age population) and Lancaster County (with 67 Hispanic voting age population, out of 11,391 total enumerated population in the 2010 census). For these experiments, we solved the optimization problem for 10,000 different instantiations of the noise. The noise here has a σ of 200, which is a high noise level selected for visualization purposes only.

Figure 6-9 shows a histogram of the distribution of post-processed populations for Fairfax County relative to the enumerated population (the 0 point on the horizontal axis represents the enumerated population of 187,160). The solid line is the theoretical Gaussian distribution. This county has a very large population of Hispanic voters, and as a result, the distribution is symmetric, and closely mirrors the Gaussian distribution.

In contrast, Figure 6-10 shows the distribution of post-processed Hispanic voting age populations for Lancaster County, with an enumerated count of 67. Because of the non-negativity constraint, the maximum negative difference between the original count and the post-processed noisy measurement is -67. All of the values in the noised distribution that would normally be below this value are pushed up to -67, hence the large spike in the histogram at this value.

6.3.4 Analytic solution to constraints

The Top-Down Algorithm and the experiments described earlier use an optimizer to find a vector that satisfies the constraints while minimizing the distance from the original values. This problem can also be solved analytically. In particular, the least-squares optimization subject to equality constraints is well known to have an analytic solution. This analytic solution involves inverting a matrix of dimension $R \times R$, where R is the number of aggregate equality constraints. Then, the non-negativity constraint, $x \geq 0$, can be satisfied iteratively by setting negative results

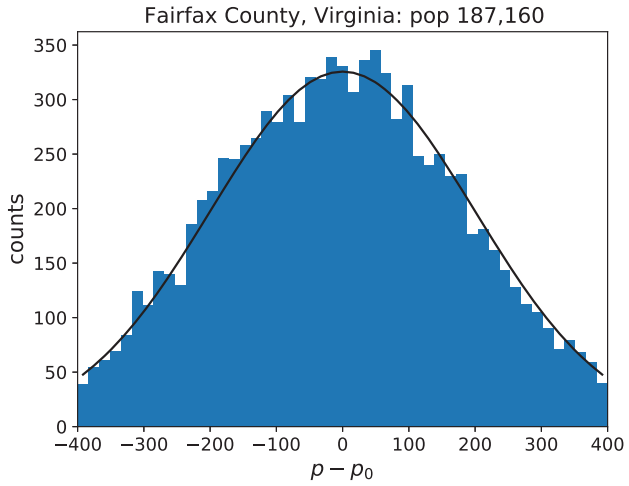

Figure 6-9: A histogram showing 10,000 trials of the post-processed counts for Hispanic over 18 population for Fairfax County, Virginia. The x-axis show population p relative to the enumerated population $p_0 = 187,160$. The noise added here has $\sigma = 200$. Since the noise is low relative to the enumerated value, the non-negativity constraint has no impact on the distribution, which matches the symmetric Gaussian distribution.

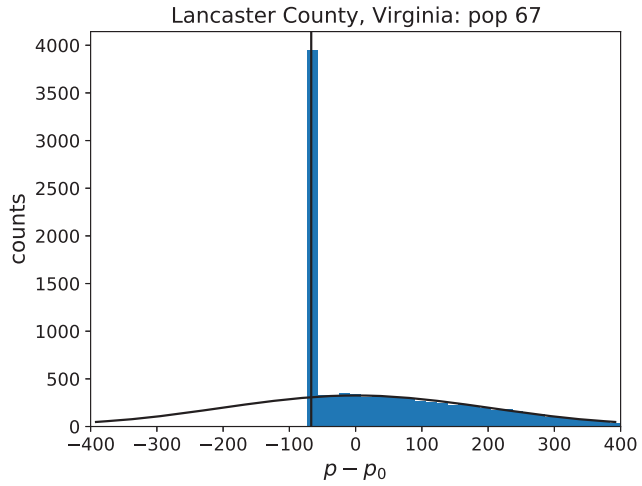

Figure 6-10: Corresponding histogram for Lancaster County, Virginia with enumerated Hispanic over 18 population of $p_0 = 67$ (note the different scales for the vertical axis). Here, the value $p - p_0$ cannot be smaller than -67, resulting in an asymmetric distribution with a large positive bias.

to zero and repeating the process. This method of solution can be faster than a numerical optimizer, and has the advantage of being simpler to explain fully and implement in a transparent and deterministic way, without needing to rely on the behavior of opaque (and possibly non-deterministic) optimizers.

Here we derive the analytic solution for the aggregate equality constraints. Using notation similar to the paper describing the Top-Down Algorithm, let $q = Q\mathbf{x}$ be the result of applying query Q to the underlying data $\mathbf{x}$. In the examples of the previous section, q are the populations of Hispanic voters enumerated for all N counties, and thus q is a vector of length $N \times 1$. Let $\mathbf{Y}$ be the noise sampled from the appropriate distribution to ensure the formal privacy property. Thus, $\tilde{M} = q + \mathbf{Y}$ is the vector of noisy measurements.

To satisfy the constraints with minimum disruption to the original results, the constraint solver seeks to find a vector $\hat{M}$ that satisfies all the constraints and minimizes the L_2 norm characterizing the difference between $\hat{M}$ and $\tilde{M}$. Thus, the function to be minimized is

$$L = \frac{1}{2}(\hat{M} - \tilde{M})^T \, W \, (\hat{M} - \tilde{M})$$

where W is a matrix of weights of dimension $N \times N$.

The R aggregate equality constraints can be represented as $B^T\hat{M} = c$ where c is a $R \times 1$ vector giving the values of the aggregate constraints and B is an $N \times R$ matrix whose columns b_m are binary vectors of dimension $N \times 1$ indicating which elements of the vector to include in the sum and the and their required total. Each of the aggregate constraints has the form $b_m^T\hat{M} = c_m$. For instance, the sum over the entire population is represented by an $N \times 1$ binary vector where each selection element is equal to one and the corresponding c_i value would be the total population.

Including these constraints in the optimization using Lagrange multipliers, the function to be optimized for becomes

$$L = \frac{1}{2}(\hat{M} - \tilde{M})^T W (\hat{M} - \tilde{M}) - \lambda_1 b_1^T \hat{M} - \lambda_2 b_2^T \hat{M} - \ \ldots \ - \lambda_R b_R^T \hat{M}$$

subject to the constraints $B^T \hat{M} = c$. The solution to this optimization is

$$W\hat{M} = W\tilde{M} + \lambda_1 b_1 + \lambda_2 b_2 \ + \ \ldots \ + \lambda_R b_R$$

which simplifies to $W\hat{M} = W\tilde{M} + B\lambda$ where λ is a $R \times 1$ vector of Lagrange multipliers. This expression can be solved for $\hat{M}$ to obtain $\hat{M} = \tilde{M} + W^{-1}B\lambda$. The vector λ can be obtained by making use of the constraint equation, $c = B^T\hat{M} = B^T\tilde{M} + B^T W^{-1} B\lambda$ from which one obtains $\lambda = \left(B^T W^{-1} B\right)^{-1}\left(c - B^T\tilde{M}\right)$. Finally, this value of λ is used to obtain $\hat{M}$,

$$\hat{M} = \tilde{M} + \left(W^{-1}B\right)\left(B^T W^{-1} B\right)^{-1}\left(c - B^T\tilde{M}\right).$$

The most time-consuming part of this computation is inverting the $R \times R$ matrix $B^T W^{-1} B$. If W is diagonal than its inverse is easy to compute. Even if W is represented by a low dimensional factor matrix, it is also a reasonably inexpensive matrix inversion. The matrix is of dimension $R \times R$ which can be inverted even for thousands of constraints.

The solution $\hat{M}$ obtained from the analytic solution using the method described above satisfies the aggregate constraints, but will likely have values that are negative, violating the non-negativity constraints. The non-negativity constraints are satisfied using an iterative procedure, which converges to the correct answer when W is diagonal. This will produce a series of $\widehat{M_i}$ solutions, until the final output $\hat{M}$ is reached. At each iteration, the largest negative value of output of the previous iteration, $\widehat{M_{i-1}}$, is identified, and that value is fixed to zero. Then, the analytic solver subroutine is then run again, but excluding the values that have been

fixed to zero. The process continues until the solver outputs a vector without any negative values. This is guaranteed to terminate since each iteration that produces any negative values reduces the size of the vector considered in the following iterations. The resulting solution, $\widehat{M_k}$, now satisfies both the aggregate constraints and the non-negativity constraints and is the final output of the analytical algorithm. The number of iterations k is at most the number of zero values in the output result.

6.3.5 Comparing the analytical algorithm with an optimizer

We conducted experiments to compare the performance of the analytical algorithm against the results from a conventional optimizer. The particular optimizer used in these experiments is the SciPy function scipy.optimize.minimize (Jones et al., 2001), as used in the experiments earlier in this section. The analytical algorithm was implemented in Python and executed on a laptop computer. For the experiments, we ran 1,000 executions of each algorithm to solve constraints of the type described in Section 6.3.3. Each optimization involved a unique sampling of the noise.

Figure 6-11 compares the difference between the optimized L_2 norms associated with the solutions obtained using these two different techniques. The solutions found are nearly identical, at the level of a part in 10^{-10}. However, where there are differences where it appears that the Scipy optimizer has stopped before converging to to the solution obtained using the analytical algorithm. We suspect this could be improved by adjusting the convergence criteria for the numerical optimizer, which would also increase its execution time.

Figure 6-12 compares the runtime of the optimizer with the runtime of the analytical algorithm. The two distributions have comparable median values, but the numerical optimizer displays tails at long times, corresponding to instances when it does not converge efficiently. Thus, there are many instances when the

analytical algorithm runs faster than the conventional numerical optimizer. Of course, none of these algorithms are running optimized, compiled code, and thus these results are to be taken lightly. The main advantages of the analytical solution are its transparency, simplicity, and repeatability.

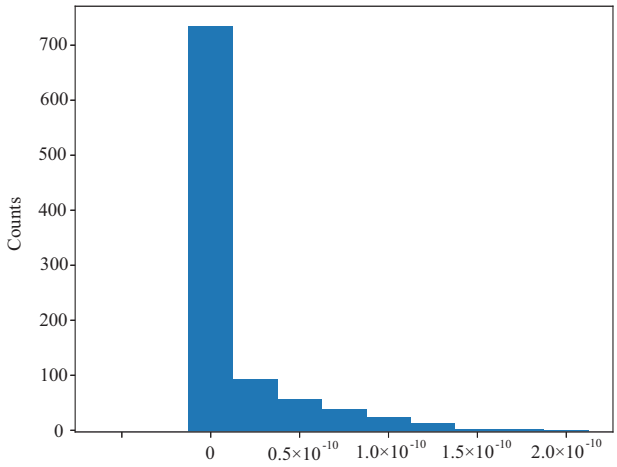

Figure 6-11: The difference between the optimized function values obtained using the SciPy optimizer and using the analytical algorithm, histogram of $\frac{f_{opt}-f_{analytical}}{f_{analytical}}$ over 1000 executions for 393 counties with 20 invariant constraints. The results are the same within one part in 10^{-10}. Where there are small differences, it appears that the SciPy optimizer stopped before converging to a solution as good as the one found by the analytical algorithm.

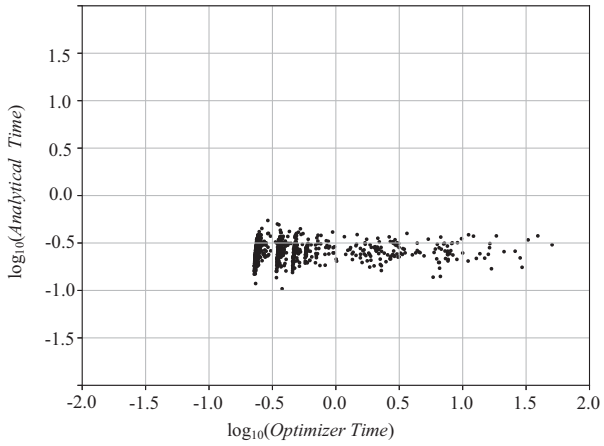

Figure 6-12: The runtime for both the SciPy optimzer and for the analytical algorithm. The two distributions have comparable median values, but the numerical optimizer displays tails at long times, corresponding to instances when it does not converge efficiently.

This Page Intentionally Left Blank

7 COMMUNICATIONS

Census data has been described by danah boyd as "democracy's data infrastructure" (2021). That lofty description captures the naturally wide and diverse range of stakeholders using the data (Section 3.1) and underscores the importance of broad and effective communications regarding any changes implemented in the collection, production, and presentation of Census data. The Census Bureau faces a particularly complex communication task in describing its changes in disclosure avoidance and their impact on the resulting data products. The diverse set of stakeholders not only have different needs and agendas for the use of the decennial tabulations, but they also have disparate levels of statistical understanding and technical capabilities.

The Census Bureau has been actively trying to communicate decisions regarding its use of differential privacy mechanisms for disclosure avoidance since at least 2018. The choice to use these privacy mechanisms and the associated communications have been met with controversy, as boyd eloquently recounts (2021):

> *When the Census Bureau decided to modernize its disclosure avoidance system . . . the bureau presumed that its data users would relish the ability to better understand the limitations underlying the statistics. The Census Bureau was wrong. Many data users—and the non-technical social groups they worked with—preferred to maintain a longstanding illusion that census data are mere counts, that published data are facts that can be distributed without accounting for uncertainty.*

At the heart of the controversy is what boyd describes as a long-held *statistical imaginary* of U.S. Census data, that "centers on an idealized vision of Census data as a near-perfect enumeration of the public that can be taken and interpreted by data users without caveats. These data can be seen as ground truth,

as the gold standard by which all other data should be measured" (boyd, d, 2021). Satisfying differential privacy requires intentionally infusing noise into released data, shattering this illusion. This has sparked conversations about noise, error, and uncertainty, challenging the prevailing statistical imaginary.

Some stakeholders understand that census data are processed, not simply collected and tabulated, and approach the data aware of mathematical analyses and statistical measures of uncertainty or error. But other stakeholders are more thoroughly committed to the statistical imaginary of Census data as a gold standard and expect the Census Bureau to provide a collection of precise and certain numbers that can serve as the basis of policy decisions without error or imperfection. These very different understandings of Census data may lead to unanticipated mismatches and challenges in communications.

7.1 Challenges and Priorities

The Census Bureau has been diligent in providing the public with extensive materials related to the 2020 Census in general, and in particular, describing changes introduced by the new disclosure avoidance mechanisms (United States Census Bureau, 2021a). The communications include blog postings, webinars, technical reports, presentations, fact sheets, and more. The Census Bureau is also planning to make its codebase transparent and available. JASON commends these efforts, and appreciates and recognizes the effort the Census Bureau has put into communicating with its stakeholders and being transparent and open as it developed plans for the 2020 census.

However, the Census Bureau has provided very little guidance on how to navigate this volume of materials given a particular stakeholder's needs and technical background. In response to recommendations from the Census Scientific Advisory Committee (2021), the Census Bureau has recognized some of these shortcomings and made commitments to continue to develop and refine their communications

and outreach. Given the breadth of stakeholder backgrounds, the additional complexity of analysis of Census data incorporating differential privacy, and the very different expertise and implicit biases of all parties, this challenge will require continual evolution and refinement. Sheer volume of communications alone will not address this challenge without adequate support for matching of resources to the needs and backgrounds of stakeholders.

Simplicity without misleading. A severe challenge in communicating complex technical ideas is to convey those ideas in a simple and clear way without misleading the intended audience. This means both avoiding statements that are technically incorrect when interpreted by a more astute reader, but also avoiding communicating in ways that may be strictly correct, but which are likely to be interpreted by a less technical audience to mean something quite different from what they should understand about what is really going on. Transparency and simplicity are essential for communicating to a diverse audience, and technical terms should be used with caution and explained at the right level when necessary. Otherwise, even well-intended communications can be more confusing than helpful.

As one small but revealing example of a specific communication JASON feels could be better framed, the following statement that is included in each 2020 Census Data Products Newsletter (United States Census Bureau, 2021c):

> *The Census Bureau is protecting 2020 Census data products with a powerful new cryptography-based disclosure avoidance system known as "differential privacy." We are committed to producing 2020 Census data products that are of the same high quality you've come to expect while protecting respondent confidentiality from emerging privacy threats in today's digital world.*

The message is nominally exciting and strictly correct, but bound to be interpreted in ways that are misleading and unhelpful. It sounds like the Census Bureau has found a new and "powerful" way of protecting census information and respondent

confidentiality. The term "cryptography-based" suggests to nearly all readers that data is encrypted in a way that makes it uninterpretable by adversaries. Technical audiences may accept that the term *cryptography* may be used to just mean that secrets are used in some way, and that differential privacy is cryptographic in that it uses secret randomness to sample the noise distribution. Some may even consider differential privacy to be "cryptography-based" because several of the researchers who developed it came from the cryptographic research community, and the paper introducing it was published in a theoretical cryptography conference (Dwork et al., 2006). But, non-technical audiences are unlikely to interpret it that way and are more likely to think that cryptography means encrypting messages so that they can only be read by a recipient with the appropriate secret key. Reading this, one can't help but get the impression that the Census Bureau wants to avoid being clear that it is using disclosure avoidance methods that involve intentionally adding *noise* to published statistics. It is understandable that the Census Bureau fears that being explicit about adding noise to results might elicit strong reactions from stakeholders. Nevertheless, it is essential to communicate that this is what the disclosure avoidance system is doing, in a manner that is easy for both technical and non-technical stakeholders to understand what this means and to start thinking about its possible implications.

The second sentence is carefully worded to express a *commitment* to continued "high quality" but misleading in hiding the fundamental tradeoffs between privacy and utility. Users need to understand that adding noise is necessary to achieve the formal privacy notion used by the Census Bureau, and that there is an inherent tradeoff between privacy and utility. While a concise announcement about disclosure avoidance methods cannot go into details, a focus on honest and direct disclosure of both the benefits and drawbacks might stave off future criticisms, or at least frame a useful starting point for discussion.

Beyond the Census Bureau's communication of their processes through the media, websites, and reports are their data products. Census data products are the most important form of communication and dissemination that the Census Bureau

does. They form the foundation for relevant, accurate, timely, reliable, and objective statistics to support the decisions of governments, businesses, individuals, households, and other organizations across the United States.

Clarity about uncertainty. The 2020 data process and resulting products are different from past decades and their release needs to support research integrity to the fullest use of their data products. As boyd notes, *"Differential privacy requires data users to contend with uncertainty, both conceptually and technically"* (boyd, d, 2021). The noise added to the Census data to satisfy formal privacy challenges the traditional "statistical imaginary" of the U.S. Census. It is therefore important to modify, rather than completely unravel the dominant statistical imaginary. One way to accomplish this would be to release the noisy measurements corresponding to the published tables (although further complicated by the microdata complexities discussed in Section 5.3 which present their own difficult communications challenges) along with the algorithms used to post-process these measurements in the production of the official tables. The Census Bureau takes strong measures to avoid ever publishing a negative or non-integral count, but including such values in the published tables would provide instant clarity to users that the statistical imaginary does not hold. It is important that users understand the biases and uncertainties that may result from the disclosure avoidance methods.

There is no question that the idea of noise and error will be challenging to many stakeholders who may not initially know what to do with the noisy measurements, and may in fact be reluctant to accept any result except an idealized, near-perfect enumeration that produces high-precision numbers. But, the Census Bureau should not use perceived lack of sophistication of some of its users as a reason to avoid providing information that will be useful to many—instead, it should aim to release its data products in a way that will make its least sophisticated users become more savvy about how they use and think about Census data. Including uncertainty estimates in decennial census data products like the PL and DHC is similar to releasing margins of errors and associated cautions for the Amer-

ican Community Survey data products. Some users ignore that information while others embrace it. Stakeholders must come to understand a common, orienting framework that acknowledges statistical error in data is a viable and robust basis for policy decisions (boyd, d, 2019). In this context, transparent communications risks seeding greater doubts initially, but the Census Bureau should recognize that clear, direct, and respectful communications with its stakeholders will overcome these doubts and strengthen trust in the long run.

7.2 Instigating Tool Support for Census Data

The use of differential privacy mechanisms introduces new statistical features into the data that will be unfamiliar to many users of previous Census data. Not only is noise added, but it is added with a distribution (e.g., discrete Gaussian) not usually taught in applied statistics courses. In the case where the noisy measurements are released with noise distribution annotations, statistical software could provide tools for users to process these data tables for maximum accuracy while providing aggregate uncertainty due to noise estimates for use in the data user's analyses. In the case where only post-processed data are available, post-processing introduces unfamiliar artifacts due to the non-negativity constraint and the least-squares fitting used to impose consistency constraints. Software tools cannot reverse the biases and inaccuracies introduced by post-processing, but may be able to estimate their impact in ways that will be useful to users.

Privacy noise is not the only driver of inaccuracy in Census data. Margins of error are already released with American Community Survey data and JASON urges that quantitative measures of uncertainty be released with all Census products. Knowing such measures allows for more accurate determination of some derived estimators and their uncertainties, but the mathematics for estimating uncertainty is not simple given the many sources of inaccuracy and their different distributions. Methods for composing uncertainties may not be familiar to Census data users.

JASON believes that the new statistical complexities are necessary in an increasingly complicated and diverse nation. One should not try to eliminate them merely for simplicity's sake—this is one argument for releasing noisy tables before post-processing adjustments, along with their post-processed counterparts. But, Census also needs to serve its less sophisticated users and make Census data accessible and useful to as many users as possible.

For some of these problems, there are well known solutions, and it is just a matter of capturing them in software tools in the right way. In other cases, the impacts of differential privacy mechanisms, and especially of post-processing adjustments, may not be estimated well by current techniques and it will be necessary to develop new methods. Census might itself undertake the theoretical work of defining best algorithms for these and other purposes and publish them as both technical papers and open source code.

Engaging Software Developers. Software tools provide a way to manage these complexities, but it is not up to the Census Bureau to produce them. Engagement with commercial and open-source organizations who develop statistical packages can help initiate third-party tool support for upcoming Census data releases, including processing noisy measurements and uncertainty measures which we hope will be made available. Popular commercial packages include SAS (SAS Institute), SPSS (IBM), Excel (Microsoft), and Stata (StataCorp). The open source software R (developed by the R Foundation, r-project.org) is also widely used. Many of these tools already provide interfaces to Census data that support users in locating, downloading, and decoding Census tables.

The Census Bureau should use its convening power to instigate discussions on the development of statistical tools and methods that can be added to the statistical packages for processing noisy measurements or for processing post-processed tables with privacy noise and post-processing uncertainties. Tools might, for example, account for the noise distribution used for each statistic and incorporate that into their estimates of statistical uncertainty. Tools could be aware

of the on-spine geographies (including the optimized block groups) and help users selected the most accurate combinations to use for a given area and to estimate its uncertainty. These estimates will help users understand the difference between using off-spine and on-spine geographical units. Tools might also provide for Monte-Carlo sampling from the probability distribution of values that could have produced a noisy measurement to help develop technical intuition about the impact of the noise.

Software tools may also be able to facilitate users in managing the differences in released data products over time, supporting users who need to make longitudinal comparisons across multiple decades. Such tools could mitigate many of the concerns about switching data products from tabulation (legacy) block groups to the optimized block groups, for example. They might also embody best practices in cases when, for privacy or other reasons, desired tables are only partly available.

By organizing workshops and giving software developers advance information about upcoming Census data releases, the Census Bureau should be able to initiate considerable efforts from the external community to develop tools for managing complex data releases, including processing noisy measurements and uncertainty estimates. The Census Bureau can also provide a central clearinghouse for work done by others. Standardizing the interfaces to such functions and providing validation tests to check implementations, would itself be valuable for human communication across platforms.

8 LOOKING TO THE FUTURE

In this section, we consider more broadly the challenges the Census Bureau faces in balancing utility and privacy in future decennial censuses and associated data products. Some of the discussion in this section applies to the long term planning for the 2030 decennial census, but other aspects may still be actionable for data products yet to be released from the 2020 decennial data.

8.1 Clarification on Title 13 Confidentiality Requirements

In its 2019 study, JASON recommended a close look at conflicts in the statutory and policy requirements for offering both absolute confidentiality to respondents and statistical accuracy in released Census products (JASON, 2020). That need has only increased in the intervening time. Where in simpler times, these conflicts have been manageable, continuing advances in technology can only exacerbate them. There is an urgent need for clarification on the interpretation of Title 13 confidentiality requirements as they apply to census data products, and perhaps even for statutory changes by Congress.

Title 13, Section 9(a) forbids the Census Bureau from making *"any publication whereby the data furnished by any particular establishment or individual under this title can be identified."* With similar intent, Title 13 Section 8(b) states that *"the Secretary may furnish copies of tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent."* Collectively, these sections of Title 13 impose strong confidentiality requirements on the data products released by the Census Bureau.

The Census Bureau's adoption of mechanisms for achieving differential privacy guarantees for the 2020 census came in large part from the reconstruction and reidentification experiments done using data from the 2010 census (described in Section 4.2). The extent to which reconstruction by itself represents an actual if,

up to now, unwitting violation of Title 13 is an arguable question of law. A lawyer representing privacy interests might argue that the data furnished by half of all respondents have been "disclosed" in the sense of Title 13. A lawyer countering these claims might argue that the individual's data have not been "identified" since the reconstruction yields no names or addresses. An additional argument could claim that many different reconstructions are possible from the same data, so reconstruction by itself does not indicate *which* records are the accurate ones. Except in a fraction of the least-populous census blocks (where new measures might in fact be needed) any "particular individual" (as in Title 13) maintains some uncertainty about any conclusions drawn about them based on reconstructed records. It seems at least arguable that, in its original legislative intent, Congress was not trying to distinguish between "disclosed" and "identified". They were not, without further explanation or definition, intending to mandate both "disclosed" and "identified" as distinct requirements, leaving room for several possible interpretations.

When the next step in the attack could lead to reidentification of an individual by linking to public data or external data sources, such as commercially available databases, there is a stronger case that Title 13 has been violated. Suppose that a respondent named P has the confidential census record (u, v, w, z) of which they view their attribute z as especially sensitive. Suppose that a reconstruction includes, among many records, an entry $(\hat{u}, \hat{v}, \hat{w}, \hat{z})$, where $\hat{\alpha}$ indicates an attribute that may or may not be an exact match for α. Now suppose that a commercial database has a record $(P, \hat{u}, \hat{v}, \hat{w})$ associated with the respondent P. These facts strongly suggest that P's private data is the reconstructed and the reconstructed record $(\hat{u}, \hat{v}, \hat{w}, \hat{z})$ can be linked with P, constituting a putative reidentification. Has this reidentification violated the confidentiality requirement? What about when there is some uncertainty such as multiple reconstructed records matching $\hat{u}, \hat{v}, \hat{w}$ with some differences in their associated $\hat{z}$ values, or some difference between the enumerated value of z and the reconstructed value $\hat{z}$? What about cases where it is still possible that an adversary who reconstructs this data can make more accurate predictions of $\hat{z}$ for P than would be possible without it?

These are difficult questions with several possible answers. As written, Title 13 deals in absolute—either an individual can be identified because of data released by the Census Bureau or they cannot. In reality, short of direct exposure of the identified microdata, things are rarely so well defined. Any information release at all that uses the individual's record in any way leaks some amount of information that potentially gets an adversary attempting a reidentification over some threshold where the inference is now harmful. But knowing whether a given data release puts an individual at risk depends on many assumptions about potential adversaries, the methods they might use, and the resources available to them. This can be tied back to the definition of differential privacy: the only way to satisfy zero privacy loss budget is if the mechanism produces the same set of outputs, with the same probability distribution, for all neighboring datasets. Since *neighboring* is transitive, this means it must always produce the same output probabilities regardless of the actual data. No matter how low the privacy loss budget is, the formal privacy guarantee cannot establish any claim that reidentification is impossible, short of not releasing any useful data at all.

Title 13 looms over everything the Census Bureau does. In particular, the Title 13 language *"whereby the data furnished by any particular establishment or individual under this title can be identified"* is often seen as sufficient reason for the alarm raised by the reconstructions demonstrated in the 2018 experiments. With a strict interpretation of Title 13, it is not necessary to carry these experiments to the next step of showing actual disclosure risk, mere identification is enough to pose a violation.

A working definition of disclosure adopted by the U.S. Federal Statistical System distinguishes three types of disclosure (National Research Council, 1993; Federal Committee on Statistical Methodology, 2005):

> "Disclosure occurs when a data subject is identified from a released file (*identity disclosure*), sensitive information about a data subject is revealed through the released file (*attribute disclosure*), or the released

> data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (*inferential disclosure*)."

In the context of the decennial census, there is no risk to an individual if it is learned that their data are included in the census. The decennial data includes all United States persons whether or not they self-report data to the Census Bureau. Attribute disclosure would occur if the collected data is compromised directly, but does not occur through published statistics. Hence, the concern for the Census Bureau should be focused on inferential disclosure risk. The risk that matters is if the released data allows an adversary to make inferences about an individual's characteristics with more accuracy and confidence than could be done without the data released by the Census Bureau.

The current wording of Title 13 provides the Census Bureau with little guidance as to how inferential disclosure risk should be balanced with the need to provide high utility data. As argued earlier, there is no way to provide any data utility without accepting some amount of inferential disclosure risk. In designing the disclosure avoidance mechanisms for the 2020 census and selecting the privacy parameters used for the production release, the Census Bureau has made the best decisions it could to balance these competing requirements. It is unclear to JASON, however, how the current interpretation of Title 13 actually guides or constrains specific decisions about disclosure avoidance.

Section 5.4 recounted the factors resulting in the selection of the privacy loss budget for the PL data release and the utility requirements that led to increases in the PL privacy loss budget. We are not aware of any factors used by the Census Bureau that would provide a limit on the maximum privacy loss budget that would adequately satisfy Title 13. If a PL privacy loss budget of, say $\epsilon = 2000$, would not have violated the confidentiality requirements, why not increase the privacy loss budget to that level to provide even better utility? Conversely, what is the basis for concluding that ($\rho = 2.63$)-zCDP satisfies the requirements of the current

interpretation of Title 13? Assuming it does, how much more privacy loss can be accepted for future Census releases derived from the same microdata before the confidentiality requirements would be violated? Answering these questions is critical in determining both what data can be included and what privacy parameters should be used for subsequent public data products derived from the decennial microdata, including the DHC release.

As an example of an alternative to the absolute requirements expressed in Title 13, many modern privacy laws explicitly recognize the complexities of balancing disclosure risks and utility goals. For example, Article 25 of the European Generate Data Protection Regulation (GDPR) privacy law states,

> *"Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall . . . implement appropriate technical and organizational measures, such as pseudonymization, which are designed to implement data-protection principles, such as data minimization, in an effective manner . . .* (Council of the European Union, European Parliament, 2016)

JASON does not necessarily favor any particular interpretation of Title 13, but believes that some new clarification of Title 13's confidentiality requirement is necessary.

8.2 Privacy Experiments

Differential privacy theorems provide a formally guaranteed upper bound on the inference risks. Correct implementations are guaranteed to not leak more information than given by this inference bound. For low enough privacy loss budgets,

such a formal guarantee is all that is needed since the inference bound given by the formal guarantee is strong enough to satisfy the desired privacy requirements. As discussed in Section 5.4, for privacy loss budgets at the levels used by the DAS the formal guarantee do not provide a reassuring inference bound. The formal privacy guarantees provided by differential privacy mechanisms at high privacy loss budgets, while valuable and important, are insufficient to provide confidence that the privacy risks are acceptable. In such scenarios, the use of formal privacy provides a principled way to select the noise distribution sampled for each query, but does not provide any comforting guarantees limiting the actual disclosure risk.

The other way to estimate actual inferential disclosure risks is to conduct empirical analyses, and JASON believes that well designed and carefully reported experiments are essential to both establishing and justifying meaningful constraints on the privacy loss budget and in effectively communicating the reasons for the disclosure avoidance mechanisms to stakeholders.

The reconstruction and reidentification experiments the Census Bureau conducted on the 2010 census data (Section 4.2) provided part of the impetus for the disclosure avoidance modernization but they have not been communicated effectively to skeptical stakeholders who are not as steeped in Title 13 as Census Bureau officials. The results from these experiments have been presented in briefings, workshops, and webinars. They have not been presented in detail in any formal Census Bureau publication. The most effective presentation that JASON is aware of is in John Abowd's supplemental declaration in the State of Alabama v. United States Department of Commerce case (Abowd, 2021), but this is not a publication that most Census stakeholders will read.

JASON recommends that the Census Bureau conduct privacy experiments in the near term to serve the following three purposes:

1. Better clarify and communicate the inference privacy risks present without noise-based privacy protections and need for modernized disclosure avoidance mitigations.

2. Measure and communicate how effectively the adopted privacy mechanisms have mitigated the risks of the demonstrated attacks.

3. Assist in evaluating the impact of design options under consideration for upcoming data products to inform decisions about these releases.

Privacy experiments typically simulate an attacker attempting to infer sensitive data from the released data and auxiliary information they may have available, and provide a measure of the effectiveness of the simulated attack. Such experiments cannot provide a guaranteed upper bound on inference risk since achieving such an upper bound would require strong assumptions that the simulated attack is the *best possible* attack that could be done. But, they are still useful. When simulated attacks demonstrate successful inferences, they provide a clear lower bound on inference risk. When carefully designed attacks fail to show any information leakage, they may provide confidence that a privacy mechanism is effective against a large class of considered adversaries.

These experiments should be designed carefully to distinguish between *distribution inference*, where the simulated adversary is learning statistical properties of the underlying distribution from the released data, and *dataset inference*, where the simulated adversary is learning something specific about individual records based on their inclusion in the data. In a scenario where the data is sampled from some actual distribution, the distinction between the two is clear — distribution inference indicates what can be inferred from the expected output resulting from any dataset that is sampled from the distribution, whereas dataset inference applies to additional inferences based on a particular dataset. What this means in the context of the census enumeration requires some consideration. The census data is not a sampling of some underlying distribution, it is meant to be a full and complete enumeration. One approach would be to imagine that the U.S. population is a sampling

from some larger underlying distribution. A simpler approach might be to just assume that any individual record should have a negligible impact on properties of the distribution. This assumption would enable direct experiments where for each execution a single individual's record is altered or removed from the dataset, and then the experiment compares what can be inferred about that individual from data resulting from executions where that individual record is modified or removed. This approach is appealing since it connects directly to the definition of differential privacy and the intuitive notion of bounding an individual's risk. Doing such experiments at the scale of the full census, however, is prohibitively expensive. Nevertheless, small scale experiments like this are possible and could produce valuable results to show the impact of varying privacy parameters. For example, such experiments could be done at the level of a single census block or tract, using the data of surrounding geographic units without modification. Representative blocks with different characteristics could be sampled for the experiments. As part of these experiments, representative individuals who are at the highest risk of exposure can be identified and used as a focus for further experimentation that could be conducted at larger scales.

A less direct, but more scalable, approach would be to provide the simulated disclosure with high quality information on the underlying distribution and to develop strong imputation adversaries based on this information. Such experiments are compelling for demonstrating the impact of census data, and may be important for considering the impact of releasing statistical information at fine granularity. These could be cases where distribution inference is a valid concern because of the detail provided in released statistics, (although these concerns are not addressed by differential privacy mechanisms). When focused solely on dataset inference, these types of experiments may overestimate the inference risks since the released data provides more information about the distribution than would be available in the simulated auxiliary information. Determining a reasonable prior for the simulated adversary may be challenging, but several options are worth considering, including priors based on data produced using a higher privacy loss budget from the same

dataset. These experiments could be useful to demonstrate the impact of privacy protections by comparing what such an adversary can infer from a version of the data released without privacy protections, with what can be inferred from data releases with varying privacy loss budgets or other adjustments to the privacy mechanism.

Finally, experiments that show relationships between inference risks, data properties, and privacy mechanisms are important and valuable. These experiments reveal how inference risks relate to an individual's characteristics, such as the size of the census block where they reside and how well represented their attributes are in the population. For some experiments, the measured inference risks will combine both distribution inference and dataset inference risks. Some experiments could be done by adjusting the privacy loss budget low enough to mostly preserve the statistical properties and then measure the difference between the inference risks without privacy protections and with the privacy protections. This would be an approximate measure of the dataset inference risks. Such experiments will be useful in understanding how inference risks relate to decisions about the detail and granularity of released data.

Next, we provide more detailed thoughts on the kinds of experiments that should be done to serve each purpose.

8.2.1 Experiments to communicate need for mitigations

The first purpose is primarily motivated by the need to effectively communicate the reasons for disclosure avoidance to stakeholders. A successful experiment of this type would demonstrate some risk that sounds bad to nearly everyone. For example, a it could show that an adversary could, with sufficient confidence to be useful, use census data to improve their predictions on the race of particular individuals or identify same sex couples with children in a particular county.

Without using results from experiments to clearly convey concrete inferential disclosure risks associated with census data releases, the Census Bureau is left needing to use abstract interpretations of Title 13 as its reason for adopting modernized disclosure avoidance methods. As an example, the most conveyed results from the Census Bureau's reconstruction and reidentification experiments on the 2010 census data are the findings that a full reconstruction at the scale of 308 million records was possible and that over 52 million people could be correctly reidentified. The first finding shows that modern computers are powerful, but provides little information on meaningful disclosure risk. Just finding one possible set of underlying microdata that is consistent with published tables does not indicate that any sensitive data has been disclosed. The second statistic about number of reidentifications sounds alarming, but does not show any inferential disclosure risk. The Census Bureau can confirm which putative reidentifications are correct from the underlying data, but an adversary would not be able to. Given the small sizes of these records without considering precision, it would be feasible to just enumerate all possible values and "correctly" reidentify everyone. Focusing on these kinds of results leaves the Census Bureau open to criticisms that it is adopting disclosure avoidance mechanisms that could result in concrete harms, with only abstract benefits. Detractors lacking understanding of the nature of the reconstructions and disclosure risks can misinterpret the results the Census Bureau has released on the reidentification experiments as doing no better than chance (Ruggles & Riper, 2021).

What matters about privacy experiments is not the average disclosure risk or even the number of individuals who may be at risk in some possibly negligible way, but how severe the risks are in the worst cases. The Census Bureau has released results, summarized in Table 8-11, from its reconstruction and reidentification experiments that make the disclosure risks meaningful and alarming, but these have not been communicated effectively to most stakeholders. The result that JASON would emphasize from these experiments is that there are 8.1 million people who live in census blocks with size less than 10, and 52% of these can be

Table 8-11: Disclosure risk demonstrated by Census Bureau's 2018 reconstruction and reidentification experiments on 2010 census data. *Unique* gives the fraction of the considered population that are have unique single-year age and sex for their census block. The *Precision* columns give the fraction of putative reidentifications that are confirmed to be correct according to the underlying CEF data. (Similar to Table 1 in Abowd (2021)).

			From Commercial Data		From CEF	
Block Size	Pop Count	Unique	Re-id Rate	Precision	Re-id Rate	Precision
<10	8.1M	0.95	0.24	0.72	0.52	**0.97**
10–49	67.6M	0.79	0.37	0.54	0.70	**0.92**
50–99	69.1M	0.59	0.44	0.42	0.75	0.82
100–249	80.0M	0.34	0.48	0.35	0.79	0.72
≥ 250	84.0M	0.08	0.50	0.27	0.85	0.60

reidentified using a reconstruction attack with perfect auxiliary data, meaning data equivalent to CEF. Of the 4.2 million people putatively reidentified this way, their sensitive attributes can be inferred with at least 97% precision.

8.2.2 Experiments to quantify inference disclosure risks from attacks

For the second purpose, the goal is to develop experiments that provide realistic estimates of how much sensitive information is disclosed by realistic inference attacks. At a minimum, the same experiments done to establish the disclosure risk for communicating the need for mitigations should be redone on the outputs resulting from the privacy mechanisms to demonstrate that the privacy mechanisms successfully thwart those attacks. These results should show a large safety margin for these attacks (for example, showing that even at a multiple of the selected privacy loss budget the attacks remain ineffective) since the attacks needed to convey disclosure risk are not designed to simulate what the best possible adversary could do.

Further risk measurement experiments may be able to provide a good approximation of an estimate for certain types of privacy risks, with some reasonable and sustainable assumptions about adversary capabilities and goals. Any claims that the experiments provide an upper limit on disclosure risk depend on assumptions about the resources available to current and future adversaries, and confidence that the simulated attack is close to the best possible attack of this type. There are potential reconstruction experiments that would be useful here. Showing that the number of possible reconstructions that match published data is high and disparate at the level of individual records demonstrates there will be a high degree of uncertainty about any record in a reconstruction that is consistent with the published data. Such experiments may not be possible at full census scale, but could still provide convincing evidence if done on some sampled geographic units.

Such reconstruction experiments could be done on the post-processed data as it would be released, but should also be done on the noisy measurements. This would add extra complications to simulating the adversary whose goal now is to find the probability distribution on the sets of underlying microdata that could have produced the released data. A sophisticated adversary would have the same goal on the post-processed tables, if that is all that were made available, but would also incorporate knowledge of the methods used in post-processing in their attack. Since it is difficult to predict how effective these methods might be, it is important to conduct inference experiments using the noisy measurements even if the Census Bureau decides not to release them.

JASON encourages the Census Bureau to conduct simulated attack experiments using simulated data. Experiments using the sensitive 2010 microdata are important for quantifying the realistic disclosure risks expected for the 2020 microdata. But this relies on the assumption that the differences in the data between the previous census and the actual data for which the privacy mechanisms will be used are small enough that it is reasonable to set privacy parameters based on experiments using the previous data. As the population of the United States becomes increasingly diverse, this assumption may not be valid. The alternative of deter-

mining the privacy parameters to use for a given data release based on the actual data used in that release would be problematic since the privacy parameters could not be set until late in the process. Further, there would be at least a theoretical risk that the determined privacy parameters, which now depend on the sensitive data, may leak some information about that data, so themselves need to be disclosed with privacy noise which would further complicate the task of users interpreting the data. An alternative would be to use synthetic data in the privacy experiments. This data would need to be generated in a way that reflects "worst case" assumptions about actual data. Various sets of synthetic data could be produced that capture different distributions about geographic units and individuals within them. In addition to giving the Census Bureau flexibility in controlling properties of data used in privacy experiments, another benefit of using synthetic data is that it could be transparently released to enable researchers outside the Census Bureau to conduct their own experiments. Judging by the academic interest in the demonstration data releases and the Census Bureau's use of differential privacy, it seems likely that if the Census Bureau publicly releases synthetic data and its own privacy experiments it would instigate considerable additional analyses from the academic research community.

Simulated attacks and synthetic data can also be used to simulate the reidentification part of experiments. Similar to the reidentification experiments done by the Census Bureau which assumed an adversary had full access to the underlying CEF data, experiments could assume an adversary with the best possible information matching exactly the underlying data used to produce the data released, but missing just the one attribute they are attempting to infer from information in the released data.

Although experiments with simulated attacks can provide useful information, it is important that any experiment like this is analyzed and presented carefully. Simulated attacks can never capture all possible methods an adversary may use to conduct an attack, including methods that may not yet have been discovered. Hence, most experiments using simulated attacks can at best provide a high level of

confidence that known attacks, and even attacks of a well-defined class of possible attacks, cannot succeed. By themselves, however, simulated attack experiments cannot provide evidence that there are no serious privacy risks.

8.2.3 Privacy experiments for informing design decisions

The final type of experiment JASON advocates for are experiments designed to inform the Census Bureau in making decisions between design alternatives, including, but not limited, to setting privacy parameters. In developing the disclosure avoidance plans for the 2020 census, the Census Bureau conducted numerous utility experiments (including the ones described in Section 6.2) to quantify the impact of different privacy mechanisms on accuracy of particular data. These experiments were very useful in informing decisions about privacy methods and parameters. Their results and the feedback from users on these experiments were influential in determining the final production PL privacy loss budget and allocation (United States Census Bureau, 2021e). Other than the initial reconstruction and reidentification experiments that resulted in the move to modernized disclosure avoidance methods, JASON is not aware of any use of privacy experiments to inform decisions about the selected disclosure avoidance methods or final parameter settings.

Setting the overall privacy loss budget for a data release is a policy decision. However, for a given privacy loss budget there are different designs that satisfy that privacy loss budget while providing different amounts of effective privacy. The corresponding perspective on utility is reflected in the experiments the Census Bureau did to adjust the geographic spine and how the privacy loss budget is allocated across queries motivated by improving accuracy for particular uses. No similar perspective appears to have been considered for privacy. The formal privacy guarantees only provide an upper bound on inference risk. As mentioned earlier, this is itself insufficient at these levels to provide an assuring guarantee. Privacy experiments should be developed and used to help inform decisions about

algorithms, privacy loss budget allocations, and data products. For example, experiments could be conducted to estimate the risks of including block-level data. It is essential to use experiments to estimate the effective risk of different kinds of inferential disclosures under different scenarios to decide among possible alternatives that satisfy the same privacy loss budget. As one concrete example, the switch from pure differential privacy to zero-concentrated differential privacy was a major change with well-understood theoretical impacts on formal privacy and was well motivated by the improvements in accuracy it enabled, but we are not aware of any disclosure experiments to evaluate its concrete impact on privacy.

JASON encourages the Census Bureau to design and maintain a suite of privacy experiments that can be executed to evaluate the impact of considered design changes on privacy. These experiments should produce quantitative outputs that give useful estimates of inference disclosure risks across a variety of adversarial assumptions. Such experiments do not need to be done at the scale of the full census, or on actual microdata, but should be efficient enough to be executed frequently to assess the likely impact of potential designs.

8.3 Planning the 2030 Census

The Census Bureau's modernization of their disclosure avoidance methods to provide formal privacy guarantees for the 2020 census has resulted in the development of novel algorithms and experiments to understand the impact of different methods of allocating privacy loss budget and mechanisms for achieving a formal privacy guarantee. Much has been learned from these efforts, including clarity on the significant challenges in meeting the competing goals of providing high quality data of the scale and comprehensiveness expected from the census while ensuring a meaningful formal privacy guarantee. The costs in terms of utility of providing a formal privacy guarantee have become more apparent, but the benefits of such a guarantee in reducing realistic disclosure risks are less clear, especially when high privacy loss budgets are used to obtain acceptable fitness for use. In particular,

it is still insufficiently clear to both Census stakeholders and to JASON as to (1) how serious the disclosure risks enabled by modern computing on census data are; and (2) whether or not the privacy mechanisms adopted for the 2020 census data products are sufficient to mitigate those risks.

Although the benefits of formal privacy guarantees are substantial in terms of communicating a clear inference bound and making strong claims about any unknown future attacks, their value may not always outweigh the cost of achieving them. The Census Bureau's move from pure differential privacy to zero concentrated differential privacy acknowledges that practical utility benefits may be a sufficient reason to weaken formal privacy requirements. Satisfying a pure differential privacy notion, or even a relaxed formal privacy notion such as the zero concentrated differential privacy definition used for the 2020 census products, is appealing to a technical community regardless of the actual privacy loss budget parameter. However, once the privacy loss budget exceeds a threshold where the inference bound is meaningful, the formal privacy guarantee by itself provides little assurance that particular attacks cannot be performed successfully.

The top part of Figure 8-1 illustrates JASON's high level perception of the process the Census Bureau used in planning the 2020 census through the PL data release. The interpretation of Title 13 combined with the results of the reconstruction and reidentification experiments led to a formal privacy requirement. The traditional expectations of how census data products are produced, along with the technical limitations of the Census Bureau's production systems, resulted in a requirement that whatever disclosure avoidance mechanisms are developed their output must be individual record microdata compatible with existent production systems. These two requirements implied a disclosure avoidance system that applies privacy-preserving noise to the results of queries on the original microdata, followed by post-processing necessary to produce the required synthetic microdata. This resulted in test and demonstration data products, which were then tested by both the Census Bureau and external stakeholders for how well they met the utility needs of particular use cases. When the trial data products were unsatis-

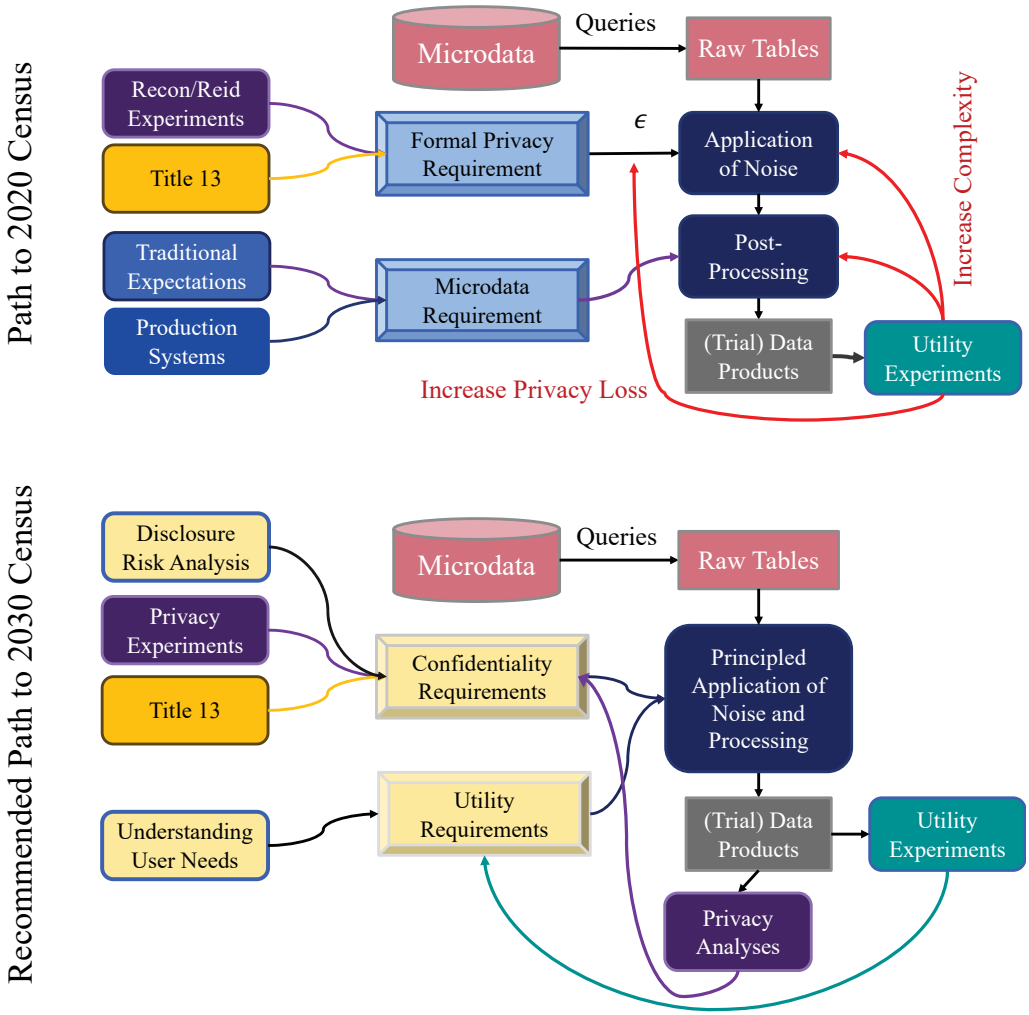

Figure 8-1: Illustration of process used to develop the 2020 census (top), and suggested process for planning for 2030 census (bottom).

factory with respect to those utility requirements, the Census Bureau revised the production process by increasing the privacy loss budget and adding additional complexity to the methods used in the DAS. This process iterated through a series of tests, resulting in the DAS and production parameters used to produce the PL data release.

The bottom half of Figure 8-1 depicts the path JASON encourages the Census Bureau to follow in planning future data products, including the 2030 census. Planning should start with as clear an understanding as possible of both the con-

fidentiality requirements and utility requirements for the data, and to formalize those requirements in a precise and transparent way. It is unlikely that all of the utility requirements can be stated precisely early in the process. The engagement with stakeholders should emphasize developing specific and careful statements of utility requirements. For the confidentiality requirements, the specific requirements should be driven by the working interpretation of Title 13 and an analysis of disclosure risks consistent with that interpretation, including the incorporation of quantifiable metrics resulting from privacy experiments (Section 8.2). The confidentiality requirements may well include formal privacy requirements. JASON recognizes the value of formal privacy both as an intrinsically valuable concept, and as a tool for developing principled solutions to practical disclosure avoidance requirements. Any formal privacy requirements adopted, however, should be motivated by an understanding of privacy semantics along with concrete scenarios to demonstrate and measure representative risks. The established confidentiality requirements should meaningfully constrain the available choices for which formal privacy definitions to use and valid ranges of their privacy loss budget parameters.

The combination of the confidentiality and utility requirements should drive the process of developing and evaluating the disclosure avoidance system and choices, following a similar open and transparent process to the one the Census Bureau followed leading up to the 2020 PL data release. But, instead of including just utility experiments bound to reduce privacy and increase complexity, the development process should also incorporate privacy analyses that include both privacy experiments and formal evaluations to evaluate how well trial data products satisfy confidentiality goals. The Census Bureau may still face difficult and challenging choices when not all requirements can be satisfied, but the resolutions to such dilemmas should be made considering information about the impact on both utility and privacy.

Finally, JASON exhorts the Census Bureau to consider complexity in all its planning decisions. Complexity has many unseen costs at the time it is introduced, including increased difficulty in communicating with stakeholders, increased im-

plementation costs and likelihood of mistakes, and additional challenges in communicating with data users. The path to the 2020 census has resulted in a method that is so complex that it is fully understood by very few (if any) people outside the Census Bureau. One of the motivations for modernizing disclosure avoidance was to avoid relying on secrecy as was necessary for the swapping-based mechanisms and to move to transparent mechanisms that can be evaluated by stakeholders. The benefits of this transparency diminish when the only way to understand the mechanisms and the decisions made about them depends on detailed understanding of internal Census Bureau systems and procedures. For the 2020 census, much of the complexity resulted from the microdata requirement imposed on the DAS. The Top-Down Algorithm also results in a great deal of complexity that was not apparent when the original decisions to adopt a top-down approach were made. Complexity tends to grow exponentially in these systems—a solution designed to solve some problem by adding a little bit of complexity, causes new problems, each of which requires new solutions sparking their own complexity spirals.

In its long-term planning efforts, JASON encourages the Census Bureau to reassess the sources of complexity that have built up in the DAS and to consider if there are ways to eliminate some of the complexities now that more is understood, or to find better solutions to some of the decisions that led to complexity spirals for the 2020 census process. In ongoing efforts, it is important to consider both short and long-term complexity costs for any decision, including the impact of the complexity on the communications challenges the Census Bureau faces and the risk that it will result in additional complications requiring further complexity.

This Page Intentionally Left Blank

9 CONCLUSIONS

The main questions posed to JASON for this study concerned whether consistency between the PL data and the DHC file was achievable and necessary. The framing and timing of the study posed some awkwardness in the relevancy of these questions. This is because as the study was being developed the final decisions on how the PL data would be published were being made. JASON was informed near the end of the in-briefings in June 2021 that the Census Bureau had decided on an approach using shared synthetic microdata to produce both the PL and DHC data. This makes achieving consistency trivial since any data products produced from the same microdata are inherently consistent. The PL data tables were released on August 12, 2021. If this synthetic microdata is augmented and used for the DHC release as is planned, then consistency will be guaranteed. However, the Census Bureau has some time yet to make that final decision and several of our recommendations encourage Census Bureau to reconsider deriving final published tables from synthetic microdata. However, JASON realizes the decisions made for the PL data release limit options Census Bureau can take for the DHC file.

JASON was asked for recommendations on how to communicate if consistency could not be achieved. Since consistency can, and will, be achieved, there is no need for any communications to specifically explain inconsistencies. JASON does make several recommendations relating to communications and they are important even when consistency is maintained. In addition, JASON was asked to advise Census Bureau as they look to 2030, and JASON provides recommendations for the Census Bureau in planning and developing future data products including the 2030 decennial census products.

We close with a consolidated list of the report's findings and recommendations.

9.1 Summary of Findings

F1 The Census Bureau has taken advantage of research advances, and their own algorithmic innovations, to control utility losses associated with achieving a formal privacy loss guarantee.

F2 The theoretical basis for the privacy methods that are used is sound. However, the effect on disclosure risk of these methods, as implemented with the selected parameters, is not well quantified.

F3 The Census Bureau sought to satisfy utility needs while minimizing formal privacy loss, but concrete disclosure risks are not sufficiently quantified to factor into decisions about disclosure avoidance options.

F4 The Census Bureau is producing the PL94-171 and DHC from generated microdata as a result of internal operational requirements. The production from microdata approach imposes otherwise unnecessary constraints that impact data quality.

F5 Census data users express concerns about data inconsistencies, but the problems associated with inconsistent data could be resolved by the Census Bureau providing guidelines for users to follow when working with inconsistent data.

F6 Block level data are not needed for the main use cases of the DHC data.

F7 Block level data (and other highly detailed data, such as age-by-year) pose the greatest reconstruction-reidentification risks.

F8 Tribal lands have different requirements including needs for highly detailed data for areas with low population.

F9 The Census Bureau has optimized its geographical hierarchy to improve the accuracy of statistics for politically important geographic areas that do not correspond to traditional on-spine entities.

F10 The detailed queries and consistency constraints enforced in producing the post-processed data are required to produce suitable microdata, as needed for the Census production system.

- The Census Bureau uses a set of detailed queries with 2,016 cells for each geographic unit (Household or Group Quarters Type [8 values] × Voting Age [2] × Hispanic/Latino origin [2] × Race [63]) down to the block level (5,892,698 populated blocks) to produce the PL94-171 data product. The statistics corresponding to these queries are not directly included in the PL data release or any future planned data releases.
- The detailed queries consume a large amount of the formal privacy budget allocated to the PL release.
- The post-processing performed to ensure non-negativity introduces bias in the results.

F11 Without access to all the noisy measurements used to produce a published value, it is difficult for users to understand how published values relate to the enumerated values.

F12 The threats and risks to both society and Census Bureau reputation from inferential disclosure attacks on Census data have not been meaningfully quantified.

F13 It is unclear if the privacy mechanisms adopted are sufficient to mitigate the vulnerabilities.

F14 The Census Bureau has put commendable effort into communicating about differential privacy and has engaged transparently with their stakeholders throughout the process of developing disclosure avoidance mechanisms for

the 2020 census but has struggled to convince stakeholders that the selected methods appropriately balance utility and privacy.

F15 The most important communications the Census Bureau does are through its public data products.

F16 Differential privacy mechanisms introduce statistical features into the data that may be unfamiliar to data users.

F17 The Census Bureau plans to take measures to avoid releasing negative population counts in upcoming data products, partly because of fears that negative values would be confusing and problematic to users. Including negative values poses communications challenges, but also provides an opportunity to clearly communicate the impact of privacy noise. Requiring non-negativity introduces bias and conceals the presence of privacy noise and complicates the methods the Census must communicate to users.

F18 Many users of Census data use widely available software tools for data analysis. Software vendors could adapt their tools to process annotated, noisy measurements and to produce more useful results from the provided data, including estimates of uncertainty.

F19 The current interpretation of Title 13's non-identification requirements is incompatible with modern technical understanding of privacy.

F20 Much has been learned about the costs and complexity of achieving formal privacy for data releases and satisfying microdata and consistency requirements, but not enough is known about whether the privacy mechanisms as implemented are sufficient to mitigate the disclosure risks that motivated adoption of formal privacy.

9.2 Summary of Recommendations

JASON's recommendations are:

R1 The Census Bureau should not prioritize consistency, either within or across data products.

R2 The Census Bureau should minimize the characteristics that are released at block level and avoid releasing DHC data at the block level.

R3 The Census Bureau should release the noisy measurements corresponding to published tables. The Census Bureau should clearly justify any use of privacy loss budget on noisy measurements that do not correspond to published statistics.

R4 The Census Bureau should release data products using the optimized block groups. These are the units with the most accurate statistics, and users should be encouraged to use the optimized block groups instead of the traditional tabulation block groups except when historical continuity is required.

R5 The Census Bureau should (i) establish standards for acceptable inferential disclosure risks, (ii) conduct experiments to understand inferential disclosure risks associated with data releases, and (iii) publish results from these experiments. JASON recommends that the Census Bureau should:

(a) Conduct experiments that make the disclosure risks concrete. For example, quantifying the ability to infer race for individuals who are of non-modal race for their block or to find cohabiting couples with children.

(b) Study the impact of privacy parameters on disclosure risk.

(c) Conduct experiments to estimate the impact of suppressing selected data such as not releasing block-level data in the DHC.

(d) Conduct experiments to simulate worst-case attacks including creative attacks that do not just perform a reconstruction followed by a reidentification, and experiments involving simulated data with high-risk properties.

R6 The Census Bureau should conduct and publish results from experiments to better understand the impact of post-processing on the accuracy and biases of computed estimates.

R7 The Census Bureau should convene meetings with tribal representatives and consider providing additional data to sovereign tribal governments in ways that satisfy their needs and recognize their distinct status.

R8 The Census Bureau should not reduce the information value of their data products solely because of fears that some stakeholders will be confused by or misuse the released data.

(a) When possible, without unduly increasing disclosure risk, all noisy measurements that are used to produce a published statistic should be released, and the process used to produce published data should be transparent and reproducible.

(b) Data releases should include explicit information on the privacy noise distribution used for each cell and any post-processing.

(c) Data releases should include estimates of all sources of uncertainty.

R9 Concurrently with releasing the noisy measurements, the Census Bureau should provide post-processed statistics, along with reproducible programs that generate the official post-processed statistics from the noisy measurements.

R10 The Census Bureau should engage with statisticians and developers of statistical software commonly used on Census data (e.g., R Consortium, Microsoft Excel, SAS, SPSS, and Stata) to develop methods for working with

annotated, noisy measurements and incorporating these into software tools.

R11 The Census Bureau should seek clarification of, or modification to, the Title 13 confidentiality requirements and plan the 2030 Census around an operational and achievable interpretation of Title 13.

R12 The Census Bureau should take an approach to the 2030 census that builds upon what has been learned from 2020, starting with developing and articulating concrete disclosure avoidance requirements for the 2030 Census data releases and designing disclosure avoidance mechanisms and data products to provide maximum utility while satisfying those requirements.

This Page Intentionally Left Blank

References

Abowd, J. (2019). Staring down the database reconstruction theorem. Presentation to AAAS Annual Meeting Feb 16, 2019.

Abowd, J. (2021). Supplemental Declaration of John Abowd for State of Alabama v. United States Department of Commerce.

Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2021). The Census TopDown algorithm. *Harvard Data Science Review* (in preparation, draft of 1 July 2021 provided to JASON).

Abowd, J. M., Stephens, B. E., & Vilhuber, L. (2005). Confidentiality protection in the Census Bureau's Quarterly Workforce Indicators. *United States Census Bureau Technical Paper TP-2006-02*. Published December, 2005. https://lehd.ces.census.gov/doc/technical_paper/tp-2006-02.pdf.

Agency for Toxic Substances and Disease Registry (2021). CDC/ATSDR Social Vulnerability Index. Published August, 2021. https://www.atsdr.cdc.gov/placeandhealth/svi/.

Allis, K. J. (2019). Letter to United States Census Bureau, representing National Congress of American Indians, July 24, 2019.

Andersson, F. (2007). Disclosure avoidance and analytical validity in "On The Map". Presented February, 2007. https://lehd.ces.census.gov/doc/workshop/2007/synthetic_otm_workshop.pdf.

Apple Computer, Inc. (2021). Differential privacy. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

Arling, G., Blaser, M., Cailas, M. D., Canar, J. R., Cooper, B., Flax-Hatch, J., Geraci, P. J., Osiecki, K. M., & Sambanis, A. (2021). A data driven approach

for prioritizing COVID-19 vaccinations in the midwestern United States. *Online Journal of Public Health Informatics*, 13(1).

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., & Sato, T. (2019). Hypothesis testing interpretations and Renyi differential privacy. *arXiv*, 1905.09982.

Bazelon, E. & Wines, M. (2021). How the Census Bureau stood up to Donald Trump's meddling. *The New York Times*. Published August 12, 2021.

boyd, d (2019). *Balancing data utility and confidentiality in the 2020 US Census*. Technical report, Data and Society, New York, NY.

boyd, d (2021). Differential perspectives: How differential privacy upended the statistical imaginaries surrounding the US Census. *Harvard Data Science Review*. Preprint August 22, 2021.

Bun, M. & Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*.

Canonne, C. L., Kamath, G., & Steinke, T. (2020). The discrete Gaussian for Differential Privacy. In *Conference on Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2004.00010.

Census Scientific Advisory Committee (2021). Recommendations and comments to the Census Bureau from the Census Scientific Advisory Committee differential privacy meeting. Published August, 2021. https://www2.census.gov/about/partners/cac/sac/meetings/2021-05/presentation-differential-privacy-recommendations-and-responses.pdf.

Cohen, A., Duchin, M., Matthews, J., & Suwal, B. (2021). Census TopDown: The impacts of differential privacy on redistricting. In *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*: Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Council of the European Union, European Parliament (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016*

(General Data Protection Regulation), Chapter IV, Article 25:Data protection by design and by default. Official Journal of the European Nation. Accessed August 30, 2021. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Dinur, I. & Nissim, K. (2003). Revealing information while preserving privacy. In *Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*.

Dong, J., Roth, A., & Su, W. J. (2019). Gaussian differential privacy. *arXiv*, 1905.02383.

Dwork, C., Greenwood, R., & King, G. (2021). There's a simple solution to the latest census fight. *Boston Globe*.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*.

Dwork, C. & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.

Dwork, C. & Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv:1603.01887*.

Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *ACM Conference on Computer and Communications Security*.

Federal Committee on Statistical Methodology (2005). Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology (second version)-2005 (SPWP22). https://www.hhs.gov/sites/default/files/spwp22.pdf.

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., & Lewis, B. (2011). A social vulnerability index for disaster management. *Journal of Homeland Security and Emergency Management*, 8(1).

Fontenot, A. E. (2019). Disclosure avoidance system design parameters and global privacy-loss budget for the 2018 end-to-end census test. 2020 Census Program Memorandum Series 2019.13. https://www2.census.gov/programs-surveys/decennial/2020/program-management/memo-series/2020-memo-2019_13.pdf.

French, C. (2014). Why demographic data matters. *Community Planning New Hampshire*. Published November, 2014. https://extension.unh.edu/sites/default/files/migrated_unmanaged_files/Resource004765_Rep6784.pdf.

Ghosh, A., Roughgarden, T., & Sundararajan, M. (2012). Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6), 1673–1693.

Gomez, J. (2021). Letter to United States Census Bureau, representing National Congress of American Indians. Written March 5, 2021.

Greenberg, A. (2016). Apple's 'differential privacy' is about collecting your data—but not your data. *Wired Magazine*. Published June, 2016. https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/.

Greenberg, A. (2017). Uber's new tool lets its staff know less about you. *Wired Magazine*. Published July, 2017. https://www.wired.com/story/uber-privacy-elastic-sensitivity/.

Gurobi Optimization, LLC (2019). Gurobi Optimizer Reference Manual.

Hansi Lo Wang and Ruth Talbot (2021). This is how the white population is actually changing based on new census data. *National Public Radio*. published August 22, 2021. https://www.npr.org/2021/08/22/1029609786/2020-census-data-results-white-population-shrinking-decline-non-hispanic-race.

Hawes, M. & Spence, M. (2021). Understanding the 2020 census disclosure avoidance system: Production settings and das accuracy metrics for the p.l.

94-171 redistricting data summary file. https://www2.census.gov/about/training-workshops/2021/2021-07-01-das-presentation.pdf.

Hotchkiss, M. & Phelan, J. (2017). *Uses of Census Bureau data in federal funds distribution: A new design for the 21st century*. United States Census Bureau.

Hughes, M. M., Wang, A., Grossman, M. K., Pun, E., Whiteman, A., Deng, L., Hallisey, E., Sharpe, J. D., Ussery, E. N., Stokley, S., et al. (2021). County-level COVID-19 vaccination coverage and social vulnerability—united states, december 14, 2020–march 1, 2021. *Morbidity and Mortality Weekly Report*, 70(12), 431.

International Statistical Institute (2003). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

JASON (2020). Formal privacy methods for the 2020 Census. (JSR-19-2F). Published April, 2020. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/plan/planning-docs/privacy-methods-2020-census.html.

Jensen, E. B., Knapp, A., King, H., Armstrong, D., Johnson, S. L., Sink, L., & Miller, E. (2020). Methodology for the 2020 Demographic Analysis estimates. United States Census Bureau. https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2020da_methodology.pdf.

Jones, E., Oliphant, T., & Peterson, P. (2001). SciPy: Open source scientific tools for Python. http://www.scipy.org.

Kairouz, P., Oh, S., & Viswanath, P. (2015). The composition theorem for differential privacy. In *International Conference on Machine Learning* (pp. 1376–1385).: PMLR.

Kirkendall, N. J., Citro, C. F., & Cork, D. L. (2020). *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*. National Academies Press (US).

LeapYear, Inc. (2021). Privacy-preserving analytics at scale. https://leapyear.io/product/.

Leclerc, P. (2019). Results from a Consolidated Database Reconstruction and Intruder Re-identification Attack on the 2010 Decennial Census. Presentation at Workshop on "Challenges and New Approaches for Protecting Privacy in Federal Statistical Programs".

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ICDE '08 (pp. 277–286). USA: IEEE Computer Society.

Mattingly, J. (2021). Quantifying Ggerrymandering A nonpartisan research group centered Duke Math. Published August 24, 2021. https://sites.duke.edu/quantifyinggerrymandering/author/0297691/.

McCarthy, R. (2004). The Bureau of Indian Affairs and the Federal Trust Obligation to American Indians. *Brigham Young University Journal of Public Law*. Published March 1, 2004. https://digitalcommons.law.byu.edu/cgi/viewcontent.cgi?article=1338&context=jpl.

McKenna, L. (2018). Disclosure avoidance techniques used for the 1970 through 2010 decennial censuses of population and housing. United States Census Bureau. Published October, 2018. https://www2.census.gov/ces/wp/2018/CES-WP-18-47.pdf.

Mervis, J. (2021). How much did a new privacy tool distort the 2020 U.S. census? Researchers ask Census Bureau for "noisy" file to measure the impact of differential privacy. *Science*. 3 September, 2021.

Mironov, I. (2012). On significance of the least significant bits for differential privacy. In *ACM Conference on Computer and Communications Security*.

Mohrman, M. & Kimpel, T. (2012). Small area estimate program: User guide. Washington State Office of Financial Management.

Mule, T. (2012a). 2010 Census coverage measurement estimation report: Summary of estimates of coverage for housing units in the united states (DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-02). https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g02.pdf.

Mule, T. (2012b). 2010 Census coverage measurement estimation report: Summary of estimates of coverage for persons in the united states (DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-01). https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g01.pdf.

National Academies of Sciences, Engineering, and Medicine (2017). *Principles and practices for a federal statistical agency*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine (2020). *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*. The National Academies Press.

National Congress of American Indians (2019). Calling Upon the U.S. Census Bureau to Consult with Tribal Nations to Ensure Both Privacy and Accuracy of Census Data for Tribal Governance. The National Congress of American Indians Resolution #ABQ-19-070. Published October, 2019.

National Research Council (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. The National Academies Press.

National Research Council (2009). *Coverage measurement in the 2010 Census*. National Academies Press.

Near, J. (2018). Differential privacy at scale: Uber and Berkeley collaboration. USENIX Enigma. Presented January 16, 2018. https://www.usenix.org/conference/enigma2018/presentation/ensign.

New York City Department of City Planning (2021). NYC Planning Community Profiles. https://communityprofiles.planning.nyc.gov/.

Oasis Labs (2021). Parcel: privacy-first data governance SDK. https://www.oasislabs.com/.

Office of Management and Budget (1997). Revisions to the standards for the classification of Federal data on race and ethnicity. Federal Register, Vol. 62, No. 210 (Thursday, 30 October 1997). https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf.

Office of Management and Budget (2006). Standards and guidelines for statistical surveys.

Privatar Labs (2021). Sharing data insights using differential privacy. https://www.privitar.com/privitar-labs/differential-privacy/.

Reamer, A. (2018). Counting for dollars 2020: the role of the decennial census in the geographic distribution of federal funds. *Initial Analysis*, 16.

Rosenberg, J. M. (2009). Defining population for one person, one vote. *Loyola Marymount University Law Review*, 42(3). Published March 1, 2019. https://digitalcommons.lmu.edu/llr/vol42/iss3/5.

Roubideaux, Y. (2021). American Indian/Alaska Native Use Cases for 2020 Census Data. Briefing to JASON representing National Congress of American Indians. Presented June 15, 2021.

Ruggles, S. & Riper, D. V. (2021). The role of chance in the Census Bureau database reconstruction experiment. *Population Research and Policy Review*. Published August 9, 2021. https://link.springer.com/content/pdf/10.1007/s11113-021-09674-3.pdf.

Salvo, J., Lobo, A., & Maurer, E. (2013). New York City population projections by age/sex & borough, 2010–2040. *Department of City Planning, City of New York, New York*.

Schafer, J. (2021). Block-level simulation of non-sampling variability in decennial census population counts. United States Census Bureau.

State of Alabama (2021). The State of Alabama, et al., v. United States Department of Commerce, et al. Lawsuit filed in United States District Court, https://www.brennancenter.org/sites/default/files/2021-03/Complaint_%202021-03-11_0.pdf.

State of Texas (2013). Shannon Perez and the United States of America v. State of Texas. Filed August 25, 2013. https://www.justice.gov/sites/default/files/crt/legacy/2013/11/19/perez_intervention.pdf.

Tang, J., Korolova, A., Bai, X., Wang, X., & Wang, X. (2017). Privacy loss in Apple's implementation of differential privacy on MacOS 10.12.

Tumult Labs (2021). Privacy protection, redefined. https://www.tmlt.io/.

United States Census Bureau (2018). Block groups for the 2020 Census – final criteria. Federal Register, 13 November 2018, https://www.federalregister.gov/documents/2018/11/13/2018-24570/block-groups-for-the-2020-census-final-criteria.

United States Census Bureau (2019a). Configuration File for E2E 2018 Test. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/E2E_2018_CONFIG.ini.

United States Census Bureau (2019b). A history of Census privacy protections. https://www2.census.gov/library/visualizations/2019/communications/history-privacy-protection.pdf.

United States Census Bureau (2020a). 2020 Census data products planning crosswalk. https:

//www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/2020-census-data-products-planning-crosswalk.xlsx.

United States Census Bureau (2020b). Standard hierarchy of census geographic entities. US Census Bureau. Published November, 2020. https://www2.census.gov/geo/pdfs/reference/geodiagram.pdf/.

United States Census Bureau (2021a). 2020 Census data products: Disclosure avoidance modernization. Accessed August 22, 2021. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html.

United States Census Bureau (2021b). Census Bureau American community survey guide for data users. Accessed August 22, 2021. https://www.census.gov/programs-surveys/acs/guidance.html.

United States Census Bureau (2021c). Census Bureau new 2020 Census data products newsletter. Accessed August 22, 2021. https://content.govdelivery.com/accounts/USCENSUS/bulletins/28c3aa6.

United States Census Bureau (2021d). Census Bureau sets key parameters to protect privacy in 2020 Census results. https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html.

United States Census Bureau (2021e). Census Bureau sets key parameters to protect privacy in 2020 Census results. Published June 9, 2021. https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html.

United States Census Bureau (2021f). Census Bureau small area estimation. Accessed on August 21, 2021. https://www.census.gov/topics/research/stat-research/expertise/small-area-est.html.

United States Census Bureau (2021g). Privacy-loss budget allocation 2021-06-08 (production settings). https://web.archive.org/web/20210701182126/https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ProductionSettings20210608/2021-06-08-privacy-loss_budgetallocation.pdf (originally published at https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ProductionSettings20210608/2021-06-08-privacy-loss_budgetallocation.pdf, but no longer available there as of 28 August 2021).

United States Census Bureau DAS Science Team (2021). (Census) alternative geographic spines. Unpublished Manuscript, provided to JASON June, 2021.

United States Code (1965). *Voting Rights Act of 1965, Pub. L. 89-110, 79 Stat. 437*. U.S.C.

United States Code (1974). *Redistricting Data P.L. 94-171*. U.S.C.

United States Code (1990). *Title 13, §§141*. U.S.C.

United States District Court for the Eastern District of Wisconsin (1982). *Wisconsin State AFL-CIO v. Elections Board, 543 F. Supp. 630*.

United States Supreme Court (1964). *Reynolds v. Sims, 377 U.S. 533*.

Wall Street Journal Editorial Board (2021). What happened to Census 'sabotage'? *The Wall Street Journal*. Published April 27, 2021.

Wang, H. L. (2021a). How 26 people in the census count helped Minnesota beat New York for a house seat. Published may 1, 2021. *National Public Radio*.

Wang, S. (2021b). Princeton Gerrymandering Project. https://gerrymander.princeton.edu/.

Wang, S., Chen, S. J., Ober, R., Grofman, B., Barnes, K., & Cervas, J. (2021). Turning communities of interest into a rigorous standard for fair districting. *Stanford Journal of Civil Rights and Civil Liberties, Forthcoming*. Published April 19, 2021.

Wasserman, L. & Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489), 375–389.

Wright, T. & Irimata, K. (2021a). Empirical study of two aspects of the TopDown algorithm output for redistricting: Reliability & Variability. Center for Statistical Research & Methodology, Research and Methodology Directorate, United States Census Bureau. Issued May 28, 2021. https://www.census.gov/content/dam/Census/library/working-papers/2021/adrm/SSS2021-01.pdf.

Wright, T. & Irimata, K. (2021b). Empirical study of two aspects of the TopDown algorithm output for redistricting: Reliability & Variability (August 5, 2021 Update). Center for Statistical Research & Methodology, Research and Methodology Directorate, United States Census Bureau. Reissued August 5, 2021. https://www.census.gov/content/dam/Census/library/working-papers/2021/adrm/SSS2021-02.pdf.