The modernization of statistical disclosure limitation at the U.S. Census Bureau

August 2020 (supersedes the 2017 version)

John M. Abowd[1], Gary L. Benedetto[2], Simson L. Garfinkel[3], Scot A. Dahl[4], Aref N. Dajani[2], Matthew Graham[5], Michael B. Hawes[2], Vishesh Karwa[6], Daniel Kifer[7], Hang Kim[8], Philip Leclerc[2], Ashwin Machanavajjhala[9], Jerome P. Reiter[10], Rolando Rodriguez[2], Ian M. Schmutte[11], William N. Sexton[12], Phyllis E. Singer[2], and Lars Vilhuber[2,12]

[1] Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau, John.Maron.Abowd@census.gov
[2] Center for Enterprise Dissemination, Disclosure Avoidance, U.S. Census Bureau, *firstname.m.lastname@census.gov*
[3] Senior Computer Scientist for Confidentiality and Data Access U.S. Census Bureau, Simson.L.Garfinkel@census.gov
[4] Economic Statistical Methods Division, U.S. Census Bureau, Scot.Alan.Dahl@census.gov
[5] Center for Economic Studies, U.S. Census Bureau, *firstname.m.lastname@census.gov*
[6] Department of Statistics, Harvard University, vkarwa@seas.harvard.edu
[7] Department of Computer Science and Engineering, Penn State University, dkifer@cse.psu.edu
[8] Department of Mathematical Sciences, University of Cincinnati, hang.kim@uc.edu
[9] Department of Computer Science, Duke University, ashwin@cs.duke.edu
[10] Department of Statistical Science, Duke University, jerry@stat.duke.edu
[11] Department of Economics, University of Georgia, schmutte@uga.edu
[12] Labor Dynamics Institute, Cornell University, {wns32,lv39}@cornell.edu

Abstract: Until recently, most U.S. Census Bureau data products used traditional statistical disclosure limitation (SDL) methods such as cell or item suppression, data swapping, input noise injection, and censoring to protect respondents' confidentiality. In response to developments in mathematics and computer science since 2003 that have significantly increased the risk of reconstruction and re-identification attacks, the Census Bureau is developing formally private SDL methods to protect its data products. These methods provide mathematically provable protection for respondent data and allow policy makers to manage the tradeoff between data accuracy and privacy protection—something previously done by technical staff. The first Census Bureau product to use formal methods for privacy protection was OnTheMap, a web-based mapping and reporting application that shows where workers are employed and where they live. Recent research for OnTheMap is implementing formal privacy guarantees for businesses to complement the existing formal protections for individuals. Research is underway to improve the disclosure limitation methods for the 2020 Census of Population and Housing, the American Community Survey, and the 2022 Economic Census. For each of these programs, we are developing new state-of-the-art privacy protection approaches based on formal mechanisms that have been vetted by the scientific community. There are many challenges in adopting formally private algorithms to datasets with high dimensionality and the attendant sparsity. In addition to formally private methods that allow senior executives to set the privacy-loss budget, our implementations will feature adjustable "sliders" for allocating the privacy-loss budget among related statistical products. The Census Bureau is implementing the settings for the privacy-loss budget and these sliders based on the decisions of the Census Bureau's Data Stewardship Executive Policy Committee.

1 Overview: Disclosure Limitation at the U.S. Census Bureau Today

The U.S. Census Bureau views disclosure limitation not just as a research interest, but as an operational imperative. The Census Bureau's hundreds of surveys and censuses of households, people, businesses, and establishments yield high quality data and derived statistics only if the Census Bureau maintains effective data stewardship and public trust.

The Census Bureau previously used traditional statistical disclosure limitation (SDL) techniques such as top- and bottom-coding, suppression, rounding, binning, noise injection, and sampling to preserve the confidentiality of respondent data. The Census Bureau is currently transitioning from these methods to modern SDL techniques based on formally private data publication mechanisms.

1.1 Legal Requirements

The Census Bureau collects confidential information from U.S. persons and businesses under the authority of Title 13 of the U.S. Code. Once collected, the confidentiality of that data is protected specifically by 13 USC §9, which prohibits:

(i) Using the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or
(ii) Making any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or
(iii) Permitting anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual records.

The privacy protections required by Title 13 are determined by the Census Bureau. Data users, including the Department of Justice and other government agencies, may be consulted regarding the criteria that determine fitness for use. Such consultation always respects the statistical-use-only requirement in the statute.

Some publications are further protected by Title 26 of the U.S. Code, which protects the federal tax information (FTI) used by the Census Bureau in the preparation of statistical products.

Confidentiality protection is intimately related to the statutory requirement that the published data be used for statistical purposes only. The definitions of "statistical purpose" and "nonstatistical purpose" were strengthened in Title III of the Foundations for Evidence-Based Policymaking Act of 2018, which is known as the Confidential Information Protection and Statistical Efficiency Act of 2018 (CIPSEA).

Additionally, the Department of Commerce (2017), in which the Census Bureau is housed, has issued directives regarding the protection of personally identifiable information (PII) and business identifiable information (BII). These directives largely mirror those issued by other government agencies and prohibit release of information

that can be used "to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc., alone or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."

1.2 Legacy methods supporting statistical disclosure limitation (SDL)

Historically, the Census Bureau has primarily used information reduction and data perturbation methods to support SDL (Lauger et al., 2014). Information reduction methods include top- and bottom-coding, suppression, rounding or binning, and sampling collected units for release in public use microdata files. Data perturbation methods include swapping, legacy noise injection systems, and partially and fully synthetic database construction. These legacy approaches start with the premise that there are specific data elements that must be protected (e.g., a person's income). A technical analyst chooses an approach from the assortment of available SDL methods that is likely to protect the data without resulting in too much damage to the published data accuracy. Usually, the selection of SDL method takes into consideration the intended uses of the published data along with assumptions about the kind of external data an intruder might have, and the types of privacy attacks an intruder might attempt.

These *ad hoc* approaches do not offer formal guarantees of data confidentiality. That is, there is no mechanism for quantifying how much privacy is being leaked from all publications based on a particular confidential database, or how one publication might interact with another publication or external data to create additional privacy risk. Furthermore, as the parameters of these legacy methods and their impact on the resulting accuracy of the data often needed to be kept confidential, there was limited opportunity for scientific scrutiny of their implementation or their effects.

1.3 Formal privacy approaches

Formal privacy methods take a different approach to protecting confidential information. Instead of starting with a list of confidential values to protect, an ad hoc collection of protection mechanisms, and ad hoc assumptions about attack models, the formal approach starts with a mathematical definition and framework for quantifying privacy risk, which permits the formulation of mathematically provable privacy guarantees against unwanted inference. Next, it implements mechanisms for publishing mathematical functions (typically called *queries*) based on the confidential data that are provably consistent with the formal privacy definition. Thus, data tables released by the statistical agency are actually modeled as a series of queries applied to the confidential data. Surrogates for public use microdata files can also be generated in this manner: instead of sampling the actual respondent data, queries are used to create formally private synthetic data. This is commonly done by first modeling the confidential data, then using the model to generate synthetic data, as discussed below.

Differential privacy (Dwork et al., 2006) is the most developed formal privacy method. It begins by specifying the structure of the confidential database to be protected, D. In

computer science, this is called the database schema; in statistics, it is referred to as the sample space. Two databases, D_1 and D_2, with the same schema are adjacent if the appropriately defined distance between them is, at most, unity. Leaving the technical details aside, say $|D_1 - D_2| \leq 1$. The universe of tables to be published from D is modeled as a set of queries on D, say Q. An element of Q, say q, is a single query on D. A randomized algorithm, A, takes as inputs D, q, and an independent random variable. The output of $A(D, q)$ is the statistic to be published, say S, which is a measureable set in the probability space defined by the independent random variable, say B. A randomized algorithm A for a publication system for releasing all of the queries in Q is ε-differentially private if, for all D_1 and D_2, with the same database schema and $|D_1 - D_2| \leq 1$, for all $q \in Q$, and for all $S \in B$:

$$\Pr[A(D_1, q) \in S] \leq e^{\varepsilon} Pr[A(D_2, q) \in S].$$

The probability is defined by the independent random variable that is used by the algorithm A, and not by the probability of observing any database D with the allowable schema (likelihood function in statistics).

There are alternative ways to define adjacent databases. For example, one method considers the databases adjacent if the record of a single person is added or removed from the database. Alternatively, the value of a single data item on a single record can be changed. Differential privacy is the mathematical formalization of the intuition that a person's privacy is protected if the statistical agency produces its outputs in a manner insensitive to the presence or absence of that person's data in the confidential database.

In differential privacy, the value ε is the measure of privacy loss or confidentiality protection. If $\varepsilon = 0$, then the two probability distributions in the definition always produce exactly the same answer from adjacent inputs—there is no difference in the output of algorithm A when given adjacent database inputs. Since the definition applies to the universe of potential inputs, and all databases adjacent to those inputs, all databases therefore produce exactly the same answer. Thus, the value $\varepsilon = 0$ guarantees no privacy loss at all (perfect confidentiality protection), but no data accuracy, since it is equivalent to releasing no data at all about the statistic S. In contrast, when $\varepsilon = \infty$, there is no confidentiality protection at all—full loss of privacy, but the statistic S is perfectly accurate (identical to what would be produced directly from the confidential input database). Thus, ε can be thought of as the *privacy-loss budget* for the publication of the queries in Q: the amount of privacy that individuals must give up in exchange for the accuracy that can be allowed in the statistical release.

Varying the privacy-loss budget allows us to move along a privacy-accuracy *Production Possibilities Frontier* (PPF) curve, as it is known in the economics literature, or along the *Receiver Operating Characteristics* (ROC) curve, as it is known in the statistics literature (Abowd and Schmutte 2019). For any attacker model, the curve constrains the aggregate disclosure risk that any confidential data might be jeopardized through any feasible reconstruction attack, given all published statistics. This budget is the worst-case limit to the inferential disclosure of any identity or item. In differential privacy,

that worst case is over all possible databases with the same schema for all individuals and items and over all external linking databases with any subset of that schema or those items.

The privacy-loss budget applies to the combination of *all* released statistics that are based on the confidential database. As a result, the formal privacy technique provides protection into the indefinite future and is not conditioned upon additional data that the attacker may have.

It is important to understand that the formal privacy protection offered by differential privacy is not absolute. Instead, it is a promise to individuals regarding the maximum amount of additional privacy loss that they may suffer as a result of a publication that is based in part on their confidential data.

To prove that a privacy-loss budget is respected, one must quantify the privacy-loss expenditure of each algorithm used to query the confidential data. The collection of the algorithms considered altogether must satisfy the privacy-loss budget. This means that the collection of algorithms used must have known composition properties.

Because the information environment is changing much faster today than when traditional SDL techniques were developed, it may no longer be reasonable to assert that a product is empirically safe given best-practice disclosure limitation prior to its release. Formal privacy models replace empirical disclosure risk assessment with designed protection. Resistance to all future attacks is a property of the design.

Differential privacy, the leading formal privacy method, is robust to background knowledge of the data, allows for sequential and parallel composability and for arbitrary post-processing edits, and enables full transparency of the implementation's source code. Differential privacy's proven guarantees hold even if external data sources are published or released later. Other formal privacy methods quantify the privacy loss that can also be mathematically established and proven, but with more constrained properties (e.g., Haney et al., 2017).

2 Expanding privacy protection for OnTheMap

Randomized response, a survey technique invented in the 1960s, was the first differentially private mechanism implemented by any statistical agency. Of course, randomized response was not recognized as being differentially private until *after* differential privacy was invented. Randomized response is sometimes called *local differential privacy.* Unfortunately, it is difficult to adapt randomized response to modern survey collection methods (Wang et al., 2016). It is the Census Bureau's experience that randomized response has a poor tradeoff between accuracy and privacy protection compared with the trusted curator model, and formal assessments of the expected additive errors of the two approaches confirm this (Kasiviswanathan et al., 2011). Vadhan notes "We have a better understanding of the local model than [multi-curator models where each trusted curator holds a portion of the confidential dataset.]

However, it still lags quite far behind our understanding of the single-curator model, for example, when we want to answer a set Q of queries (as opposed to a single query)." (Vadhan 2017)

The first production application of a formally private disclosure limitation system by any organization was the Census Bureau's OnTheMap (residential side only), a geographic query response system for studying residence and workplace patterns.

The Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES), the data used by OnTheMap, is a partially synthetic dataset that describes geographic patterns of jobs by their employment locations and residential locations as well as the connections between the two locations (U.S. Census Bureau, 2016). A job is counted if a worker is employed with positive earnings during the reference quarter and in the quarter prior to the reference quarter. These data and marginal summaries are tabulated by several categorical variables. The origin-destination (OD) matrix is made available by ten different "labor market segments". The area characteristics (AC) data–summary margins by residence block and workplace block–contain additional variables including age, earnings, and industry. The blocks are defined in terms of 2010 Census blocks, defined for the 2010 Census of Population and Housing. The input database is a linked employer-employee database, and statistics on the workplaces (Quarterly Workforce Indicators: QWI) are protected using noise injection together with primary suppression (Abowd et al., 2009, 2012).

For OnTheMap and the underlying LODES data, the protection of the residential addresses is independent of the protection of workplaces. Protection of worker information is achieved using a formal privacy model (Machanavajjhala et al., 2008); work is in progress to protect workplaces using formal privacy as well (Haney et al., 2017).

3 SDL methods supporting the 2020 Census of Population and Housing

The 2000 and 2010 Censuses of Population and Housing applied SDL in the form of record swapping, but this fact was not always obvious to data users. The actual swapping rate was kept confidential, as was the overall impact that swapping had on data accuracy (McKenna 2018).

The Census Bureau successfully tested the feasibility of producing differentially private tabulations of the redistricting data (PL94-171) for the 2018 End-to-End Census Test, and is currently in the final stages of algorithm development, for the full-scale implementation of differentially private protections for the 2020 Census of Population and Housing.

In October 2019 the Census Bureau re-released data from the 2010 Census using an early prototype for the 2020 Census Disclosure Avoidance System (DAS) (U.S. Census Bureau 2019). Called the 2010 Demonstration Data Products, this system was the subject of a December 2019 meeting of the Committee on National Statistics, where

attendees compared the statistical accuracy of these data products with previous data publications based on the 2010 Census. The source code used to prototype the 2010 Demonstration Data Products was released the following month. This code base included 33,853 lines of Python programs and 1263 lines of configuration files. In July 2020, the Census Bureau subsequently re-released the 2010 Census data protected using an updated version of the 2020 Census DAS, as the 2010 Demonstration Privacy-Protected Microdata File 2020-05-27 (U.S. Census Bureau 2020).

The differentially private mechanisms designed for the 2020 Census support the following products:

- **Public Law (PL) 94-171** files for redistricting;
- **Demographic Profiles and Demographic and Housing Characteristics files** for demographic statistics pertaining to individuals and housing units;
- **Detailed tabulations on race, ethnicity, and household composition;**
- **Privacy Protected Microdata,** the actual microdata from which published data products were tabulated; and
- **Noisy Measurements,** the actual differentially private statistics used to create the consistent microdata, to allow researchers outside the Census Bureau to produce independent statistical products without suffering the unavoidable accuracy loss that results from the post-processing of the differentially private statistics to convert them back into microdata for tabulation.

The Census Bureau has designed its differentially private algorithms to allow a selected number of queries based on the confidential data to be reported exactly. Such queries are called *invariants.* The Census Bureau currently plans the following invariants for the 2020 Census data publications:

- Total number of people by state;
- Total number of housing units (aggregate of occupied and vacant housing units) by block; and
- Total number of group quarters within three-digit group quarters type by block. Group quarters types are defined in Table P43 (U.S. Census Bureau 2012).[1]

While the inclusion of these invariants requires clarification of the formal privacy guarantees under differential privacy, they were considered necessary to permit public scrutiny of the state apportionment totals, and to permit the public-input component of the Local Update of Census Addresses (LUCA) program.

[1] Table P43, "Group Quarters Population by Sex and Age by Group Quarters Type," is in Segment 6 of the 2010 Census SF1. It can be downloaded from https://www2.census.gov/census_2010/04-Summary_File_1/.

Key disclosure limitation challenges include:

1. Ensuring consistency across tables by respecting the invariants enumerated above;
2. Producing block-level microdata for use by the Census Bureau's tabulation system to support production of traditional data products;
3. As was true of historical systems like swapping, there is difficulty detecting coding errors, particularly as they relate to verifying privacy-loss guarantees;
4. Determining how much of the privacy-loss budget should be spent per household; e.g., whether it should be proportional to household size;
5. A lack of high-quality usage data from which to infer relative importance of data products; and
6. The lack of public input data with which to develop and simulate the mechanism.

Key policy-related challenges include:

1. Communicating the global disclosure risk-data accuracy tradeoff effectively to the Data Stewardship Executive Policy Committee (DSEP) so that they can set the privacy-loss budget and the relative accuracy of different publications,
2. Providing effective summaries of the social benefits of privacy vs. data accuracy, so that DSEP, in particular, can understand how the public views these choices.

Throughout each decade, the Census Bureau also conducts special tabulations of small geographic areas such as towns. Those tabulations also impact privacy, and they also undergo SDL.

4 SDL methods supporting the American Community Survey (ACS)

The American Community Survey (ACS) is the successor to the long form survey of the Census of Population and Housing. The housing unit survey includes housing, household, and person-level demographic questions about a broad range of topics. There is a separate questionnaire for those residing in group quarters. The Census Bureau sends this survey to approximately 3.5 million housing units and group quarters each year and receives approximately 2.5 million responses. Weighted adjustments account for nonresponse, in-person interview subsampling, and controlling to pre-specified population totals. The ACS sample is usually selected at the tract level and is designed to allow reliable inferences for small geographic areas and for subpopulations, when cumulated across five years. ACS sampling rates vary across tracts. On average, a tract will have approximately thirty-five housing units and ninety people in the returned sample.

The Census Bureau releases one-year and five-year ACS data products. Five-year tables are released either by block group or by tract. One-year tables have been released only

for geographies containing at least 65,000 people. A recent Census Bureau Disclosure Review Board (DRB) decision allowed some one-year tables to be released for areas of at least 20,000, due to the termination of the three-year data products. The Census Bureau also releases one-year and five-year Public-Use Microdata Sample (PUMS) files for both persons and housing units. These PUMS contain samples of ACS microdata records (1% and 5% samples, respectively) with geographic detail limited to Public Use Microdata Areas (PUMA). PUMAs are special non-overlapping areas that partition each state into contiguous geographic units containing roughly 100,000 people.

The feasibility of developing formally private protection mechanisms given current methodological and computational constraints, the large number of ACS variables, and the desire for small area estimates is undemonstrated. The Census Bureau is actively pursuing this research, seeking to leverage advances from other data products. The Census Bureau is also funding cooperative agreement opportunities for research into the use of formal privacy for surveys in general. As an intermediate step to provide additional privacy to ACS respondents, the Census Bureau is experimenting with the development of non-formally private synthetic data using statistical and machine learning models to replace the current SDL methods.

Key disclosure avoidance challenges include:

1. **High dimensionality:** there are roughly two hundred topical module variables with mixed continuous and categorical values,
2. **Geography**, with estimates needed at the Census tract and block-group levels,
3. **Variable associations** across people in the same household,
4. **Outliers** in the economic variables,
5. **Survey weights** due to sampling, nonresponse, and population controls.

These challenges stem from high dimensionality combined with small sample sizes. Small geographies and sub-populations are important for data users, even if they do not always properly incorporate the sampling uncertainty when using these data. Tract-level and even block group-level data are critical for many applications, including the ballot language determinations in Section 203 of the Voting Rights Act. In addition to legislative districts, tabulations for many special geographies published by the Census Bureau, including cities and school districts, are built from smaller component geographies.

The large margins of error for small geographies allow some scope for introducing error from SDL without significantly increasing total survey error. Modelling can introduce some bias in exchange for massive decreases in variances by borrowing strength from correlations.

The research team is currently developing methods to protect ACS microdata utilizing synthesis models combined with a validation system. The overall approach is:

1. Build a chain of models, simulating each variable successively given the previous synthesized variables (Raghunathan et al., 2001). Currently, the team is assessing the use of classification trees for this purpose (Reiter, 2005);
2. Create synthetic microdata from these models for all records and all variables, creating fully synthetic data; and
3. Allow users to validate results from the synthetic microdata against the internal data. Validated results would have to meet the same standards for disclosure avoidance as all other public data releases and would be limited in quantity to statistics required for the stated purpose.

As opposed to current ACS Public Use Microdata Samples (PUMS), this fully synthetic microdata would not use internal files that have already had SDL applied to them as its source; rather, the ACS program will generate an Internal Reference File (IRF) to serve as the source. The IRF can serve as a baseline dataset for assessing survey accuracy without the confounding impacts of SDL methods, and will allow the research team to evaluate the effects of synthesis on privacy and accuracy in isolation.

The research team is considering other models for protecting tabular output, including hierarchical and spatio-temporal models.

Validation servers, verification servers[2], and access to the Federal Statistical Research Data Centers (FSRDCs) may be the solution for research questions for which the modernized SDL approach leads to reasonable uncertainty regarding the suitability of published data for a particular use. An advantage of the formally private methods being tested for both the 2020 Census and the ACS is that they permit quantification of the error contributed by the SDL; hence, the inferences drawn from these data can be corrected for the impact of the uncertainty added to protect privacy. Their suitability for use in a particular application can also be assessed without reference to the confidential data. This property of modernized SDL provides a means for applying objective criteria to a researcher's claim that the published data are suitable or unsuitable for a particular use.

5 SDL research supporting the 2022 Economic Census

Every five years the Census Bureau sends survey forms to nearly four million U.S. business establishments, broadly representative of all geographic regions and most private industries, to conduct the Economic Census. The Economic Census is based on a complete enumeration for certain types of businesses, and sampling of other, mostly smaller, businesses. The Census Bureau defines an *establishment* as a specific economic activity conducted at a specific location, and asks companies to file separate reports for

[2] Validation servers provide the data user with the results of their query calculated on the internal data with SDL performed on the result. Verification servers provide the data user with some measure of how confident they should be with the result of their query calculated on the synthetic data.

different locations and when multiple lines of activity are present at the same location. The Economic Census survey collects information from sampled establishments on the revenue obtained from product sales in the industries in which they operate, as well as information on employment, payroll, and other establishment characteristics.

Key policy challenges include:

1. Specifying the entity to be protected: multi-unit companies operate many establishments with different forms. From a legal standpoint, it is not entirely clear which entity (company, establishment, or something else) must be protected.
2. Defining what constitutes sufficient protection. Requirements to protect fact-of-filing may imply that whether a given business appears must be protected. However, it may not be necessary to protect certain business attributes that are in the public domain.

Key disclosure avoidance challenges include:

1. **Outliers** in the economic variables and generally high skewness;
2. **Sparsity** of data in cells disaggregated down to the North American Industry Classification System (NAICS) subsector and county level;
3. **Hybrid** sampling and enumeration design combined with an edit and imputation stage that complicate privacy models;
4. **Associations** among economic variables that increase disclosure risk; and
5. **Complex** publication schedules that require consistency over time and efficient allocation of privacy-loss budgets across releases.

The Census Bureau's disclosure modernization efforts for the Economic Census have followed two potentially complementary paths. Beginning in 2017, an interdisciplinary team at the Census Bureau partnered with academic colleagues to evaluate the feasibility of developing synthetic industry-level microdata. The methods under consideration are not formally private, but would allow publication of more detailed information while maintaining disclosure protections comparable to the cell suppression methods currently in use. Kim, Reiter, and Karr (2016) present methods of developing synthetic data on historic Economic Census data from the manufacturing sector. An inter-divisional team has applied two synthetic data models to 42 industries from the 2012 Economic Census covering eighteen economic sectors. Input data were limited to full-year reporter businesses (births, deaths, and seasonal businesses were excluded). The synthetic data were evaluated for fidelity in summary tabulations of items collected for all sectors. The team is currently evaluating the disclosure risk for these approaches. Kim and

Thompson are working on a separate synthetic data model that includes businesses that are part-year reporters.

In 2020 an additional team began work to develop formally private disclosure avoidance methods appropriate to economic data in general, and the Economic Census in particular. Since the publication schedule does not require release of microdata, the team is exploring modifications of the differential privacy paradigm that could be directly applied to tabular summaries and yield provable privacy guarantees. Specifically, they are considering a variant of the model developed in Haney et al., (2017) as well as other approaches in the smooth sensitivity framework (e.g. Nissim, Raskhodnikova and Smith, 2007). The sparsity of the published tables may require a modification of these methods to ensure consistency and data quality while keeping privacy loss at acceptable levels. The team intends to develop methods applicable to the County Business Patterns and Economic Census First Look products, which have relatively simple structure. From there it will hopefully be possible to adapt those methods to more complex Economic Census products.

6 Challenges and meetings those challenges

In differential privacy, the commonly used flattened histogram representation of the universe is calculated as the Cartesian product of all potential combinations of responses for all variables. This representation is often orders of magnitude larger than the total population even when structural zeroes (impossible combinations of values of variables, such as grandmothers who are three years of age) are imposed. One promising approach is approximate differential privacy, where the limiting factor depends only on the logarithm of the inverse probability of algorithmic failure.

Policy makers, including the Census Bureau's DSEP, must have enough information about the privacy-loss/data accuracy trade-off to make an informed decision about ε, and its allocation to different tabular summaries. In some cases, the chosen amount of noise injection from differential privacy may limit the suitability for use of the published statistics to more narrowly defined domains than has historically been the case.

The strategy for producing the tabular summaries is to supply the official tabulation software with formally private synthetic data that reproduce all of the protected tabulations specified in the redistricting and summary file requirements. In generating high quality synthetic microdata, one needs to consider integer counts, non-negativity, unprotected counts (e.g., total state population), and structural zeroes.

To execute this approach, the Census Bureau needs generic methods that will work on a broader range of datasets. In addition, it may be difficult to find meaningful correlations that are not represented in the model. To address this, the model must anticipate the types of analyses that data users might wish to conduct. As a result, better model-building tools are needed, as well as generic tools for correlating arbitrary models

with the ones used to build the synthetic data. Ongoing engagement with data users is also essential to help identify these intended uses of the published data.

Reproducible-science methods will be required to use synthetic data effectively.

Data are often collected with a complex sample design with considerable missing data and in panels of longitudinal data. Research is ongoing to ensure that weighted, longitudinal analysis using differentially private data will continue to produce "good results and good science" to the data users.

7 Approaches to gauge data accuracy and usefulness

There are multiple methods to assess data accuracy, also known as analytical (or inference) validity. Machanavajjhala et al. (2008) conducted experiments comparing differentially private synthetic data to the actual data for OnTheMap. They saw value in coarsening the domain to limit the number of "strange fictitious commuting patterns." Karr et al. (2006) and Drechsler (2011) advocate calculating confidence interval overlaps for parameters of interest, whether univariate, bivariate, or multivariate.

There is value in calculating all such metrics described above for parameter estimates calculated from:

- non-perturbed data (exact counts) where we expect parity; and
- parameter estimates that were not captured in the joint distributions modeled in the synthetic data, where one would not expect to uncover comparable results.

Disclosure limitation is a technology. It shows the relationship between privacy loss, which is considered a public "bad", and data accuracy, which is considered a public "good". A differentially private system can publish extremely disclosive data. This happens if the privacy-loss budget is set very high. The extremely disclosive data will likely be very accurate. That is, inferences based on these data will be nearly identical to those based on the confidential data. But extremely disclosive, albeit formally private, data also permit a very accurate reconstruction of the confidential data relative to the reconstruction possible with smaller privacy-loss budgets.

The teams at the Census Bureau working on formal privacy methods for statistical disclosure limitation have been charged by DSEP with developing technologies with adjustable parameters to control the privacy loss and data accuracy during implementation. Those technologies will be summarized with a variety of supporting materials. The Disclosure Review Board will make a recommendation regarding the appropriate formal privacy technology and parameter settings, including the privacy-loss parameter ε. The Data Stewardship Executive Policy Committee will review that recommendation and make the final determination. The published data will implement the recommendations of DSEP. Although more explicit than in previous censuses, this is the same chain of recommendation and approval that was used in 2000 and 2010.

This transition to innovation involves significant retooling of methods for the Census Bureau's career mathematical statisticians, computer scientists, subject matter experts, project and process managers, and internal stakeholders. This transition will help the Census Bureau lead similar innovation across the U.S. Federal Government and beyond.

8 References

Abowd, John M. and Ian M. Schmutte "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices," American Economic Review, Vol. 109, No. 1 (January 2019):171-202, DOI:10.1257/aer.20170627.

Abowd, John M., R. Kaj Gittings, Kevin L. McKinney, Bryce Stephens, Lars Vilhuber, and Simon D. Woodcock (2012). *Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures for Related Time Series.* 12-13. U.S. Census Bureau, Center for Economic Studies.

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon D Woodcock (2009). *The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators.* In Producer Dynamics: New Evidence from Microdata, edited by Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. University of Chicago Press.

Department of Commerce, Office of Privacy and Open Government (2017). *Safeguarding Information.* http://osec.doc.gov/opog/privacy/pii_bii.html#PII

Drechsler, Jörg (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation.* New York: Springer.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006) Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third conference on Theory of Cryptography* (TCC'06), Shai Halevi and Tal Rabin (Eds.). Springer-Verlag, Berlin, Heidelberg, 265-284. DOI=http://dx.doi.org/10.1007/11681878_14

Garfinkel, Simson, John M. Abowd, and Christian Martindale, Understanding Database Reconstruction Attacks on Public Data, Communications of the ACM, February 2019.

Garfinkel, Simson, John M. Abowd, Sarah Powazek, Issues Encountered Deploying Differential Privacy, Workshop on Privacy in the Electronic Society, Toronto, Canada - October 15, 2018.

Haney, Samuel, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber (2017). *Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics*, SIGMOD'17, May 14-19, 2017, Chicago, Illinois, USA, DOI: 10.1145/3035918.3035940.

Karr, A.F., C.N. Kohnen, A. Oganian, J.P. Reiter, and A.P. Sanil (2006). *A framework for evaluating the utility of data altered to protect confidentiality.* The American Statistician 60, 224-232.

Kasiviswanathan, Shiva Prasad, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith (2011). *What can we learn privately?.* SIAM Journal on Computing 40, no. 3: 793-826.

Kim, Hang J., Jerome P. Reiter, and Alan F. Karr (2016). *Simultaneously Edit-Imputation and Disclosure Limitation for Business Establishment Data.* Journal of Applied Statistics online: 1-20.

Lauger, Amy, Billy Wisniewski, and Laura McKenna (2014). *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research.* Research Report Series (Disclosure Avoidance #2014-02). Washington: Center for Disclosure Avoidance Research, U.S. Census Bureau.

McKenna, Laura (2018). *Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing.* Working Papers 18-47, Washington: Center for Economic Studies, U.S. Census Bureau.

Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber (2008). *Privacy: Theory Meets Practice on the Map.* Proceedings: International Conference on Data Engineering. Washington, DC, USA: IEEE Computer Society, 277-286.

Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk, and Peter Solenberger (2001). *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models.* Survey Methodology 27(1). Citeseer: 85-96.

U.S. Census Bureau (2012). 2010 Census Summary File 1: 2010 Census of Population of Housing. September 2012. U.S. Census Bureau. https://www.census.gov/prod/cen2010/doc/sf1.pdf

U.S. Census Bureau (2016). OnTheMap: Data Overview (LODES Version 7). U.S. Census Bureau. https://lehd.ces.census.gov/doc/help/onthemap/OnTheMapDataOverview.pdf

U.S. Census Bureau (2019). 2010 Demonstration Data Product. October 2019. U.S. Census Bureau. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html

U.S. Census Bureau (2020). 2010 Demonstration Privacy-Protected Microdata File 2020-05-27. July 2020. U.S. Census Bureau. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf/?#

Vadhan, Salil (2017). *The Complexity of Differential Privacy.* March 14, 2017. https://privacytools.seas.harvard.edu/files/privacytools/files/complexityprivacy_1_01.pdf

Vilhuber, Lars and Ian M. Schmutte (2016). *Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy.* http://digitalcommons.ilr.cornell.edu/ldi/33/

Wang, Yue, Xintao Wu, and Donghui Hu (2016). *Using Randomized Response for Differentail Privacy Preserving Data Collection.* Workshop proceedings of the EDBT/ICDT 2016 Joint Conference. March 15, 2016, Bordeaux, France. http://ceur-ws.org/Vol-1558/paper35.pdf

9 Disclaimer

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.