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

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

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

Abstract: Most U.S. Census Bureau data products use traditional statistical disclosure limitation (SDL) methods such as cell or item suppression, data swapping, input noise infusion, and censoring to protect respondents' confidentiality. In response to developments in mathematics and computer science since 2003, the Census Bureau is developing formally private SDL methods to protect its data products. These methods will provide mathematically provable protection to respondents and allow policy makers to manage the tradeoff between data accuracy and privacy protection—something previously done by technical staff. The Census Bureau's OnTheMap tool is a web-based mapping and reporting application that shows where workers are employed and where they live. OnTheMap was the first production deployment of formally private SDL in the world. Recent research for OnTheMap has incorporated 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 2017 Economic Census. For each of these programs, we are developing models to create synthetic microdata, from which we can create aggregated estimates. There are many challenges in adopting formally private algorithms to datasets with high dimensionality and the attendant sparsity. We are also developing approaches for gauging the synthetic data's accuracy and usefulness for specific applications. In addition to formally private methods that allow senior executives to set the privacy-loss budget, our implementation will feature adjustable "sliders" for allocating the privacy-loss budget among related sets of tabular summaries. The U.S. Census Bureau will implement the settings for the privacy-loss budget and these sliders using recommendations from the Data Stewardship Executive Policy Committee, as was done in the 2000 and 2010 Censuses.

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 Bureau's hundreds of surveys and censuses of households, people, businesses, and establishments yield high quality data and derived statistics only if the Bureau maintains effective data stewardship and public trust.

The Bureau has traditionally used statistical disclosure limitation (SDL) techniques such as top- and bottom-coding, suppression, rounding, binning, noise-infusion, and sampling to reserve the confidentiality of respondent data. The Bureau is currently transitioning from these SDL 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.

Some publications are further protected by Title 26 of the U.S. Code, which also protects the federal tax information (FTI) used by the Bureau in the preparation of statistical products, primarily from businesses. Additionally, the Department of Commerce (2017), in which the 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 Current methods supporting statistical disclosure limitation (SDL)

Currently, the Bureau primarily uses information reduction and data perturbation methods to support SDL (Lauger et al., 2014). Information reduction methods include swapping, top- and bottom-coding, suppression, rounding or binning, and sampling collected units for release in public use microdata files. Current data perturbation methods include swapping, noise infusion, and partially and fully synthetic database construction. The current approach starts with the premise that there are specific data elements that must be protected (e.g., a person's income). A technical analyst choses 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.

These ad hoc approaches do not offer formal guarantees of data confidentiality. That is, a person's income may be suppressed in a cell, but it may be possible to reconstruct that person's income by combining information published elsewhere within the statistical tables; that is, without using any external data.

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 and an ad hoc collection of protection mechanisms, the formal approach starts with a mathematical definition of privacy. Next, it implements mechanisms for publishing *queries* based on the confidential data that are provably consistent the formal privacy definition. Thus, the tables released by statistical agency are actually modeled as a series of queries applied to the confidential data. Surrogates for public use microdata samples (PUMS) files can also be generated in this manner: instead of sampling the actual respondent data, the queries are used to create formally private synthetic data. This is 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 the sample space. Two databases, D_1 and D_2, with the same schema are neighbors if the appropriately defined distance between them is unity. Leaving the technical details aside, say $|D_1 - D_2| =1$. 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 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 $|D_1 - D_2| =1$, 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 persons 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 neighboring 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 neighbors of 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 encrypting the statistic S. In contrast, when $\varepsilon = \infty$, there is no confidentiality protection at all—full loss of privacy, but the statistics S are 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 2017). The curve constrains the aggregate disclosure risk that confidential data might be jeopardized through any feasible reconstruction attack, given all published statistics for any attacker. 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.

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.

To prove that a privacy-loss budget is respected, one must quantify the privacy-loss expenditure of each publication or published query. 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 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 allows for arbitrary post-processing edits. 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 (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, although it was not a conscious decision, and the technique is difficult to adapt to modern survey collection methods (Wang et al., 2016).

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 infusion 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 Census Bureau is testing the feasibility of producing differentially private tabulations of the redistricting data (PL94-171) for the 2018 End-to-End Test. It is currently in the process of algorithm development and obtaining the cloud computing environment necessary to scale the research to the requirements of the Census of Population and Housing. For the full 2020 Census, the Bureau will extend the methods used for the 2018 End-to-End test to the tabulations in Summary File 1.

The differentially private tabulations for the 2020 Census will support the following products:

- **Public Law (PL) 94-171** for redistricting,
- **Census Summary File (SF) 1** for demographic and housing counts, and
- **Geographical Hierarchy** from the national to the block level, exploiting parallel composition to efficiently use the privacy-loss budget.

By agreement with the Department of Justice (2000), the Census Bureau will provide exact counts at the Census block level for the following variables:

- Number of people: total, age 18+ (voting age), and less than age 18,
- Number of vacant housing units, and
- Number of householders, which is equal to the number of occupied housing units.

Key disclosure limitation challenges include:

1. Ensuring consistency by respecting the unaltered counts enumerated above,
2. Respecting joins; e.g., to group people into households,
3. Large memory/time requirements for explicitly stored universes and well-understood low-dimensional approximations,
4. Difficulty detecting coding errors, particularly as pertains to verifying privacy-loss guarantees,
5. Communicating analytical results clearly to, and in a format useful for, policy makers,
6. A lack of high-quality usage data from which to infer relative importance of data products, and
7. Determining how much of the privacy-loss budget should be spent per household; e.g., whether it should be proportional to household size.

The 2010 and 2000 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 is confidential, as is the impact that swapping had on overall accuracy. 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.

Key policy-related challenges include:

1. Communicating the global disclosure risk-data accuracy tradeoff effectively to the Disclosure Review Board and Data Stewardship Executive Policy Committee 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 DESP, in particular, can understand how the public views these choices.

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 Bureau sends this survey to approximately 3.5 million housing units each year and receives approximately 2.5 million responses. Weighted adjustments account for nonresponse, in-person interview subsampling, and raking to pre-specified population controls. 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 averaged 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 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 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 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 Bureau is actively pursuing this research, seeking to leverage advances from other data products. As an intermediate step, the Bureau is experimenting with non-formally private synthetic data using statistical 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 level,
3. **Preserving associations** among variables across people in the same household,
4. **Outliers** in the economic variables,
5. **Dealing with weighting** due to sampling and nonresponse adjustment.

These challenges stem from high dimensionality combined with small sample sizes. Small geographies and sub-populations are important for data users. 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, many special geographies published by the Census Bureau, including cities and school districts, are dependent upon small 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 for massive decreases in variances by borrowing strength from correlations.

The research team is considering the following approach:

1. Build a chain of models, simulating each variable successively given the previous synthesized variables (Raghunathan et al., 2001)
2. Build a formally private version of these models, if feasible.
3. Create microdata samples from these models.
4. Create tables from these microdata samples.

Validation servers, verification servers, and access to the 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 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 from the published data are correct. Their suitability for use in a particular application can, therefore, be assessed without reference to the confidential data. This property of the modernized SDL provides a means for applying objective criteria to a researcher's claim that the published data are unsuitable for a particular use.

5 SDL research supporting the 2017 Economic Census

Every five years the Bureau sends survey forms to nearly four million U.S. business establishments, broadly representative of the complete U.S. geography and most private industries, to conduct the Economic Census.

The Bureau defines an *establishment* as a specific economic activity conducted at a specific location. The Bureau asks companies to file separate reports when operating at different locations and when multiple lines of activity are present at a given location. The Economic Census is thus a mixture of a complete enumeration for certain types of businesses, and sampling of other types.

The Economic Census collects information from sampled establishments on the revenue obtained from product sales ("products") in any given industry. Establishments can report values from a wide variety of potential products. The reported product values are expected to sum to the total receipts reported earlier in the questionnaire. Often, product descriptions are quite detailed, and many products are mutually exclusive. Consequently, legitimate missing values occur frequently. Good predictors such as administrative data and other survey data are available for variables such as revenue, payroll, and employment, but auxiliary data are not available for the other items.

The key challenge that the development team will focus on is the disclosure limitation process for North American Product Classification System (NAPCS) product estimates that are new to 2017. The current plan is to release product and product-by-industry tabulations that satisfy predetermined privacy and reliability constraints and to release supplemental synthetic industry-level microdata files, pending the outcome of the research discussed below.

Beginning in 2017, an interdisciplinary team at the Census Bureau partnered with academic colleagues to evaluate the feasibility of developing synthetic industry-level microdata comprising general statistics items and selected products. Specific products may differ by industry and the level of model estimation (industry, industry by state) will need to be determined in the course of the research.

Kim, Reiter, and Karr (2016) present methods of developing synthetic data on historic Economic Census data from the manufacturing sector. The goal is to extend their multivariate joint model to accommodate additional Economic Census industries, modifying them as the research indicates. There are other publications already approved for the 2017 Economic Census; hence, the synthetic data must satisfy additional constraints—specifically the preservation of published margins. The proposed methods allow for multiple imputation variance estimation. It has not been determined whether the multiple imputation variance estimates for the synthetic data will need to approximately match the published variance estimates.

In addition to developing usable datasets, there is an additional goal of teaching users to use synthetic data to produce their own tabulations and conduct their own analyses. The team thus needs to consider usage and analysis by outside users.

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 three years of age) are imposed.

Policy makers, including the Data Stewardship Executive Policy Committee of the Census Bureau, 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 infusion from differential privacy may limit the suitability for use of the published statistics to more narrowly defined domains than has historically been true.

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., voting age population), and structural zeroes.

To execute this approach, the 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 analysis that many downstream users might 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.

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 establishing 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.
- parameters 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 are also very accurate. That is, inferences based on these data are 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 forward its recommendation to the Director. The published data will implement the recommendations of DSEP, as they have for the past two decennial censuses. 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, IT specialists, 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 Schmutte (2017). *Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods*. Under review. http://digitalcommons.ilr.cornell.edu/ldi/37/.

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

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.

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.

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