1-24-2017

# Proceedings from the 2016 NSF–Sloan Workshop on Practical Privacy

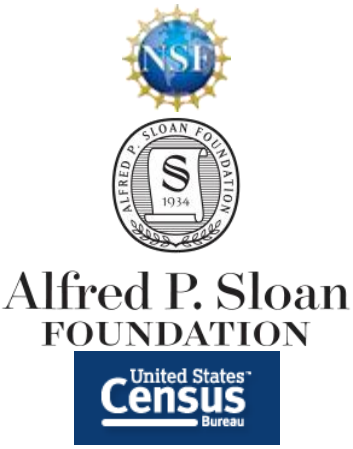
Lars Vilhuber
*Cornell University ILR School*, lars.vilhuber@cornell.edu

Ian M. Schmutte
*University of Georgia*, schmutte@uga.edu

# Proceedings from the 2016 NSF–Sloan Workshop on Practical Privacy

**Abstract**
On October 14, 2016, we hosted a workshop that brought together economists, survey statisticians, and computer scientists with expertise in the field of privacy preserving methods: Census Bureau staff working on implementing cutting-edge methods in the Bureau's flagship public-use products mingled with academic researchers from a variety of universities. The four products discussed as part of the workshop were 1. the American Community Survey (ACS); 2. Longitudinal Employer-Household Data (LEHD), in particular the LEHD Origin-Destination Employment Statistics (LODES); the 3. 2020 Decennial Census; and the 4. 2017 Economic Census. The goal of the workshop was to 1. Discuss the specific challenges that have arisen in ongoing efforts to apply formal privacy models to Census data products by drawing together expertise of academic and governmental researchers 2. Produce short written memos that summarize concrete suggestions for practical applications to specific Census Bureau priority areas.

**Comments**
Funding for the workshop was provided by the National Science Foundation (CNS-1012593) and the Alfred P. Sloan Foundation. Organizational support was provided by the Research and Methodology Directorate at the U.S. Census Bureau and the Labor Dynamics Institute at Cornell University.




# Proceedings from the NSF–Sloan Workshop on Practical Privacy

Held on Friday October 14, 2016 in Washington DC

Lars Vilhuber and Ian M. Schmutte, *Editors*

Funding for the workshop was provided by the National Science Foundation (CNS-1012593) and the Alfred P. Sloan Foundation. Organizational support was provided by the Research and Methodology Directorate at the U.S. Census Bureau and the Labor Dynamics Institute at Cornell University.

# Table of Contents






Funding through NSF CNS-1012593 and Sloan Foundation.

# Disclaimer

Many of the participants of this workshop are employees or contractors of the U.S. Census Bureau. The opinions, discussions, and conclusions reported in these proceedings are those of the participants and do not necessarily represent the views of the U.S. Census Bureau, the National Science Foundation, or the Alfred P. Sloan Foundation. This document has not undergone the review accorded Census Bureau publications and no endorsement should be inferred. While many participants contributed their notes to the summaries, all final editing was done by the editors. All results have been reviewed to ensure that no confidential information is disclosed.

# Goals and Methods of the Workshop

*Lars Vilhuber and Ian M. Schmutte*

Protecting the privacy of respondents, whether they may be individual people or companies, while at the same time providing useful and meaningful statistics, are two competing goals to which statistical agencies devote considerable effort. With the recent surge in new methods, triggered in part by the growing interest in differential privacy (Dwork et al. 2006), old ways of "doing things" are being questioned, and yet new methods are almost never drop-in replacements. Even when the theory has been fully mapped out, and algorithms abound (Dwork and Roth 2014), translating theory into functional, repeatable, scalable, and accepted practices for providing real data is rare (Machanavajjhala et al. 2008).

On October 14, 2016, we hosted a workshop that brought together economists, survey statisticians, and computer scientists with expertise in the field of privacy preserving methods: Census Bureau staff working on implementing cutting-edge methods in the Bureau's flagship public-use products mingled with academic researchers from a variety of universities. The four products discussed as part of the workshop were

1. the American Community Survey (ACS);
2. Longitudinal Employer-Household Data (LEHD), in particular the LEHD Origin-Destination Employment Statistics (LODES); the
3. 2020 Decennial Census; and the
4. 2017 Economic Census.

The goal of the workshop was to

1. Discuss the specific challenges that have arisen in ongoing efforts to apply formal privacy models to Census data products by drawing together expertise of academic and governmental researchers
2. Produce short written memos that summarize concrete suggestions for practical applications to specific Census Bureau priority areas.

We formed two break-out groups in each of two sessions. Each group had 2 hours to discuss the issues surrounding a particular data product. Allocation to each group was not random: at least one Census Bureau product or team lead provided an overview of the characteristics of the product, the issues being faced in regards to disclosure avoidance, and ongoing efforts to address them. The entire group was free to discuss any aspect of theory, implementation, etc. No conclusion needed to be reached.

A consensus summary of the discussion was compiled by note takers in the group, and was presented to the plenum midway through the workshop, and before the end of the workshop. The note


Funding through NSF CNS-1012593 and Sloan Foundation.

takers then drafted a summary, which was subsequently circulated among the group members for review and correction. The final summary appears in these proceedings.

## Common Threads

A few common threads appear throughout the discussions on new disclosure avoidance techniques. First, the technological challenges: **existing data structures** differ widely. The difficulties in applying new disclosure protection mechanisms are compounded by the variety of data being collected, from the language spoken in a household as reported by a person (American Community Survey), to the presumed location of a worker's economic activity as reported by an employer (LEHD), to the quantity and cost of consumed polyester for the production of broadwoven fabrics (one part of the Economic Census). **Existing data consumers**, the stakeholders in the data production exercise, expect continuity inthe quality and diversity of statistics that they have obtained in the past. **Legal requirements** restrict where disclosure avoidance can bite. For instance, the U.S. Congress has dictated that the population counts (in total, and for age groups over and under 18) derived from the Decennial Census must be accurate at the state and national levels. As of this writing, Freedom of Information Act (FOIA) requests on government workers can constrain the amount of privacy protection that can be afforded to individuals who at some point were in the employment of the U.S. federal government in tables generated from the LEHD infrastructure.

These constraints on how the data are produced, who they are produced for, and the legal framework surrounding those issues, are non-trivial complications of an already quite complicated endeavor.

Another significant commonality across all groups is the expressed **need to measure uses** of the data. Historically, many tabulations have been created in a comprehensive and exhaustive measure, yet many of those tabulations may never be used. A better assessment of which (past and potential future) queries are of primary interest is a key need for all groups.

Finally, it was noted in all groups that a particular challenge is the **protection of hierarchically structured** data. Whether it be workers within employers (LEHD), or individuals as parts of households (ACS and Decennial Census), or products within firms (Economic Census), this feature of many social datasets is absent from much of the newer literature on formally private systems, and is considered a key challenge to be addressed by the scientific community.

## Next steps

The members of the various teams have expressed an interest to continue discussing these topics in these workshops. A follow-on workshop is planned for the Spring of 2017.

# References

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: Halevi S., Rabin T. (eds) Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, vol 3876. Springer, Berlin, Heidelberg. doi:10.1007/11681878_14.

Dwork, Cynthia, and Aaron Roth. 2014. “The Algorithmic Foundations of Differential Privacy.” Foundations and Trends in Theoretical Computer Science. now publishers, Inc., 211–407.

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. International Conference on Data Engineering. Washington, DC, USA: IEEE Computer Society, 277–86.

# American Community Survey

*Based on notes prepared by Amy Lauger, William Sexton, and Lars Vilhuber*

## Summary of Data Product

The American Community Survey (ACS) is the successor to the prior long form of the Decennial Census of Population and Housing. The housing unit survey includes 24 housing and household questions and 48 person-level demographic questions about a broad range of topics. There is a separate questionnaire for those residing in group quarters (U.S. Census Bureau 2014). ACS produces about 11 billion tabular summaries. The survey is sent to approximately 3.5 million housing units each year and we collect about 2.5 million responses after nonresponse, etc. Weighting adjustments are made to account for nonresponse, in-person interview subsampling, and raking to 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. Tracts are designed to have a population of around 4000 people, and ACS sampling rates vary across tracts. However, on average, a tract will have approximately 35 housing units and 90 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 while historically one-year tables have been released only for geographies containing at least 65,000 people. A recent decision has allowed some tables to be released for areas of at least 20,000, due to the termination of the three-year data products.

From web site tracking, the ACS staff have some data indicating demand for the tables. Selected economic characteristics and ACS demographic tables are requested over two million times per year. However, the majority of public tables are never requested.

Currently, the main disclosure avoidance methods for the official estimates include data swapping for the household population and data synthesis for the group quarters population. Some noise is added in certain cases (Lauger, Wisniewski, and McKenna 2014). A major drawback of data swapping is that there is no known way of quantifying its privacy guarantee.

In addition to the tables, the Census Bureau releases Public Use Microdata Samples (PUMS)[1]. Additional disclosure avoidance methods for these files currently include:

1. Limiting geographic detail to areas with a population of 100,000
2. Categorical variable coarsening
3. Topcoding

[1] See also http://www.census.gov/programs-surveys/acs/technical-documentation.html

The current research into improved disclosure avoidance (DA) techniques, led by Jerry Reiter (Duke University), is considering both synthetic data and formally private data.

## Summary of Discussion

The team is concerned about the feasibility of developing formally private protection mechanisms given current methodological and computational constraints and the large number of ACS variables. The creation of synthetic data is a backup option that would allow the replacement the current DA methods, which don't allow for any provable quantification of privacy protection or transparency about the effects on data quality.

It was noted in the discussion that it is not the research team's job to set the privacy budget but rather to describe the utility/privacy frontier to the decision makers.

The main challenges were described as:

1. High dimensionality: around 200 topical variables
2. Geography
3. Within-household relationships
4. Outliers in economic variables
5. Protect tables directly or create microdata and produce tables from there?
6. One-year vs five-year data products?

The main disclosure challenges stem from high-dimensionality combined with small sample sizes. Small geographies and sub-populations are important for key uses of the ACS: Tract-level data, and even block group-level data are critical for many data users. These data have very high error rates. However, users often use these geographies as building blocks to create local areas according to their own salient definitions, that in aggregate have more acceptable error rates. Also, many special geographies published by the Census Bureau, including cities, school districts, etc., are dependent upon tracts and block groups.

Given the small sample sizes at the tract level, the team assumes that tract may be the lowest level feasible for modeling, and even that might be tricky. However, many 5- year ACS estimates are published at the block group level. The team hasn't determined yet if and how estimates at that level can be provided.

The team is collecting metrics on which data are requested most often. Those metrics, and collaboration with the analysts in other directorates, will help inform which relationships are most important to maintain.

Editing and imputation is used to clean the data. The specs for the edit and imputation processes are thousands of pages long. The team believes that incorporating editing and imputation into the DA systems is infeasible at this stage. For now, the team will start with the edited data, and will consider if edit must be rerun post-DA.

The large margins of error for small geographies allow for some scope for introducing error from DA without significantly increasing total survey error. Modeling can introduce some bias for massive decreases in variance by borrowing strength from correlations. We need to think about all of this from a total survey error perspective, where the error includes that introduced from sampling, edits, nonresponse, disclosure, etc.

John Abowd indicated that Census Bureau officials are highly concerned about PUMS files right now. The ACS team will need to consider phasing in improvements as they are ready instead of waiting to implement all changes at once.

The research team is currently considering the following approach:

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

John Abowd commented that this approach does not take advantage of correlations in tables.

Preventing p-hacking is an added bonus of differential privacy methods. We wish to provide researchers with enough information to make valid inferences but can't support all possible analysis due to budget limitations. It will be necessary to explicitly acknowledge whose data interests are most favored. This is part of the previously mentioned transparency.

Maintaining within-household relationships was briefly mentioned this as a problem. Examples given:

1. A biological daughter cannot be older than her father
2. Within household clustering, for example, by race

The group did not discuss solutions.

Jerry Reiter mentioned concern for **outliers** in variables such as property value and income. However, there is some protection from extreme outliers in that the ACS only collects data up to a certain level of precision. We should model outliers independent of the real data. The RDCs will be the solution for research questions for which this leads to unacceptable data usefulness.

The group briefly discussed the role of **weights**. Survey weights can themselves add disclosure risk. The current plan is to build a synthetic dataset that replicates the ACS sample, not the population. Then weights will be created.

The group only briefly discussed time levels for aggregation and protection. Options are either creation of 5-year releases with overlap, or 1-year released that are then combined to 5-year averages afterwards. Overlap between 5-year tables may result in strange temporal bias issues.

Chris Clifton suggested that publishing this list of challenges is in itself a good exercise. It may encourage other academics to work on these problems. The group suggested various journals (Challenge, Chance, Science) and various differential privacy workshops.

The group indicated that outside researchers cannot work on the problem without access to the right data and discussed whether the PUMS would be a good starting point. The PUMS files could be altered to simulate various aspects like smaller geographies and outliers, etc. Some believed this would be a feasible place to start while others felt they were not, given that they are already privacy protected. Others suggested a completely simulated microdata file assuming uniform distributions.

Jerry Reiter summarized that the PUMS or some variation of the PUMS favors inferences about data usefulness while a simulated dataset favors inferences about privacy.

John Abowd suggested that the ACS team prioritize figuring out how to model data nationally in an accurate and parsimonious way and then determine how to get to lower areas. The team should determine where it does well and doesn't do well. He also suggested that an ideal tool would be a tunable workload assessor, so a stakeholder can decide what aspects are most important, while keeping algorithm development and policy decisions separate. Order matters with privacy budgets. John Abowd gave the example of Decennial Census redistricting files being published first, and the resulting need to ensure that those stakeholders don't eat up the privacy budget before other stakeholders get their relevant files. Dan Kifer suggested giving each stakeholder a certain epsilon and it's up to him to decide how to use it. When they've exhausted it, they will need to rely on data other stakeholders chose to get.

## References


Lauger, Amy, Billy Wisniewski, and Laura McKenna. 2014. "Disclosure Avoidance Techniques at the US Census Bureau: Current Practices and Research." Research Report Series (Disclosure Avoidance) no. 2014-02. Washington: Center for Disclosure Avoidance Research, US Census Bureau. https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf.

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. 2014. "American Community Survey - Design and Methodology." 2.0. U.S. Census Bureau. http://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html




Funding through NSF CNS-1012593 and Sloan Foundation.

# LEHD Origin-Destination Employment Statistics (LODES)

*Based on notes prepared by Matthew Graham, Vishesh Karwa, and Lars Vilhuber*

## Summary of Data Product and Context

LEHD Origin-Destination Employment Statistics (LODES) 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 as well as 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 10 different “labor market segments.” The area characteristic (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 Decennial Census. The input data is a linked employer-employee dataset, and statistics on the workplaces (Quarterly Workforce Indicators, QWI) are protected using noise infusion together with primary suppression (Abowd et al. 2009; Abowd et al. 2012). For OnTheMap and the underlying LODES data, the protection of the residential addresses of the workers is achieved using a formal privacy model (Machanavajjhala et al. 2008). The two protection mechanisms are independent of each other. The data is hierarchical (workers work for employers), but can be visualized as stemming from an evolving bipartite graph.

The current challenge is to build a formal privacy model that protects both residences and workplaces.

## Summary of Discussion

The discussion started with a brief discussion about the need to protect. Why are businesses (in LODES, and other Census products) protected as they are? Since U.S.C. Title 13 is not very precise, there is room for interpretation, and the question revolved around the regulatory and historical reasons for the type of protection mechanism chosen. The answer was primarily that historical precedent guides current practice. Business X's location and type (industry classification) are observable and not subject to protection, but characteristics of the operation of the business, and of its employees, are protected. On the residential side, all characteristics, including location, are protected.

Part of the discussion revolved around the intrinsic longitudinal aspect of the underlying data, and its effect on both protection and publication. The current QWI publications protect cross-sectional characteristics of firms' operations, but do not protect longitudinal aspects thereof (e.g., job creation). LODES statistics are based on a single cross-section each year, are not revised, but draw new noise

every year. This potentially adds a lot of noise, and may be inefficient. However, current research has not found a satisfactory answer, and it is viewed as an open research question how to properly do longitudinal protection.

Much time was spent on attempting to understand the query workload, aka the public's need for data. All possible tabulations are produced, and protected, potentially expending privacy budget (in a vague sense) without corresponding use. What parts of the overall tabulations are actually used by data users? Could a reduced set be produced, discarding unused tabulations, and achieving better accuracy in released tabulations with the same overall protection ("privacy budget"). The core data are available as flat CSV files available for anonymous download, yielding it nearly impossible to assess actual usage of all tabulations. However, the Census Bureau has started measuring usage in its user-friendly download interfaces: LED Extraction Tool[2], QWI Explorer[3] (both for QWI), and OnTheMap[4]. Census personnel noted that quite a lot of queries stem from automated tools that use non-traditional geographic (free-form) queries, which cannot be classified by traditional Census geography. These are driven by tools like OnTheMap for Emergency Management[5], though they do not necessarily derive from user needs. Researchers have in the past attempted to analyze the query logs, without great success so far, but it is seen as a critical and important open task. It was noted, however, that past queries are not necessarily a good indicator of future queries, and any mechanism that takes into account past queries must also have a mechanism to accommodate future unanticipated queries. A side question briefly explored to what extent query logs are themselves confidential.

Other than the simple usage statistics, the question was asked how utility was measured for LODES. The three metrics (L1 error, errors in ranking queries, and fractional error) were outlined (Machanavajjhala et al. 2008; Haney et al. 2017).

A related question referred to the reporting of margins of error. Generically, input data error (non-survey error) are known to swamp protection errors (unpublished statistics by Abowd et al). It was noted that this might reduce the need for protection if the system were integrated, which it is not. The error from all imputation methods (unit-to-worker imputation, imputation of demographic characteristics, imputation of firm characteristics, see Abowd et al., 2009) is not taken into account when computing the protection. It was noted however that most of the imputation models used are multiple imputation models, and composition (the ability to combine the imputation and protection mechanisms) is difficult and an unsolved problem, with much of the theoretical work still missing.

One feature that was highlighted as an efficient use of the privacy budget. In some rare cases, data were revised (due to improvements of the underlying data), and it was feasible to resynthesize only the

[2] https://ledextract.ces.census.gov/
[3] https://qwiexplorer.ces.census.gov/
[4] https://onthemap.ces.census.gov/
[5] https://onthemap.ces.census.gov/em/


Funding through NSF CNS-1012593 and Sloan Foundation.

affected (small) area, and a new $\varepsilon$ computed only for that area. This does compose, and the impact on the overall privacy budget can be computed, and is small.

## Challenges For the Future

1. Computing and reporting of margin of error
2. Improved mechanisms for updating Longitudinal Employer-Household Dynamics (LEHD) Program data
3. Revising the protection mechanism for residential data, and integrating with workplace protection
4. Adapting protection mechanism to workload

## References


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. doi:10.2139/ssrn.2159800.

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 Micro Data*, edited by Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. University of Chicago Press.

Haney, Samuel, Ashwin Machanavajjhala, John Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. 2017. "Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics." In *SIGMOD 2017 (accepted)*.

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. International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 277–86.

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




Funding through NSF CNS-1012593 and Sloan Foundation.

# 2020 Decennial Census

*Based on notes prepared by William Sexton, Phil Leclerc, and Lars Vilhuber*

## Summary of data product and context

The 2020 Decennial Census is the next constitutionally mandated count of the US population. An exact count of the population is required for the proper apportionment of the US House of Representatives. Data from the census is also used in the distribution of federal funds to local communities. The census consists of 10 questions, and requires the enumeration of all persons living in a given domicile, collecting data on demographic characteristics (age, sex, ethnicity, race), and information about ownership of the domicile.[6]

Once collected, summaries of the raw data are produced. Summary File 1 (SF1) contains population tables, housing tables, and geographic tables of the data collected from the Decennial Census (U.S. Census Bureau 2012). The Decennial Census yields tables for each of the 50 states, District of Columbia, Puerto Rico, and the United States. The Public Law 94-171 (PL94) Census Redistricting Data Summary File is a particularly important subset of SF1. Congress has dictated that the population counts (total, under 18, over 18) in these tables must be accurate at the state and national levels. For missing data, Bureau of Commerce policy dictates that the Decennial Census use hot-deck imputation. Any privacy measures must respect these stipulations.

The discussion focused on research efforts underway to implement a formally private confidentiality protection system. The goal of this system is to produce publicly usable, simulated microdata sets (PUMS), and to provide formally private, accurate answers to the "most important" subset(s) of the Summary File tables[7], with particular attention paid to the PL94 tabulations.

## Summary of discussion

Key challenges that were highlighted include:

1. ensuring consistency
2. respecting joins and answering nonlinear queries
3. large memory/time requirements for explicitly stored universes, and well-understood low-dimensional approximations may either approximate poorly or complicate update rules
4. difficulty detecting coding errors, particularly as pertains to verifying privacy guarantees
5. communicating analytic results clearly to and in a format useful for policy makers
6. a lack of high-quality usage data from which to infer relative priority of data products

[6] For more information, see http://www.census.gov/programs-surveys/decennial-census/2020-census/about.html.

[7] Note that it is not clear *a priori* whether separate budgets should be used for the PUMS and tabular answers, or if the PUMS should only reflect the information used to construct the tabular answers.

7. determining how much of the privacy budget (ε) should be spent per household (e.g. proportional to house size?)

The team conducting this research is led by Dan Kifer. They are experimenting, similarly to the ACS team, with the Multiplicative Weights Exponential Mechanism (MWEM) (Dwork and Roth 2014; Hardt, Ligett, and McSherry 2012), which produces differentially private tabular summaries of underlying microdata.

Some participants expect, based on prior experience with MWEM, that the formally private system may produce less accurate data than the traditional statistical disclosure limitation techniques (e.g. swapping, additive/multiplicative noise, cell suppression) currently in use to generate the tables in SF1. Others pointed out that due to the lack of formal description of the extent of protection in current systems, such as swapping, make that conclusion less robust. Reduced data quality is a 'hard sell' for the Census Bureau's policy leadership. If formally private data is systematically inconsistent, heavily biased, or sufficiently noisy in ways that break down frequently used statistical machinery, many stakeholders in the use of Census Bureau data will be unhappy.

A particular challenge is the hierarchical structure of the Decennial data: Households are sampled, and within households, data on individuals is collected. These two levels must remain consistent even in the protected data: for instance, in the synthesized data, children must not be older than the parents, within the same household. Thus data consistency is a major concern in the implementation of formally private Decennial Census data releases.[8]

The group acknowledged the complexity of the Census Bureau's process for cleaning the raw data to generate the gold standard files, and that this process is in principle affects the privacy system. At present the research team is not formally addressing this issue (they are treating the gold standard files as "truth"). However, in this context, Kobbi Nissim mentioned that a technique known as "privacy amplification" may be useful in using sampling variance to gain extra privacy 'for free' in a PUMS.

The group also discussed criteria by which to evaluate the quality of formally private data relative to that produced using Census' traditional SDL methods. To provide basic benchmarks, we can use either the simple, workload-based Laplace mechanism or a uniformly distributed PUMS to provide "sanity checks" on the quality of more sophisticated methods. More sophisticated comparisons are also possible; we might, for example, seek to compare and contrast a differentially private 2020 Census with a *k*-anonymity-protected 2020 Census, although comparisons of this kind are fraught with technical issues (e.g., how do we choose ε and *k* to standardize this comparison appropriately?). Use of the PL94 Summary Files to conduct redistricting suggests that some reasonable, automated metrics might be generated by considering the impact of a formally private mechanism on districts generated using standard district-determination models/algorithms from the political science literature. We may also

[8] The geography-level records in the public use files (block, tract, etc.) pose an entirely different challenge: being in significant part computed directly from the raw data, they are currently considered out-of-scope for the formal privacy project, but their release procedure is not formally private.

consider comparing and contrasting formally private data products from the 2020 Census to data from the 2010 Census and to recent iterations of the American Community Survey, since changes relative to these earlier data products will be important for many common uses of 2020 Census data. Lastly, differential privacy might be motivated from an axiomatic framework, possibly rooted in analysis of relevant Constitutional or Congressional mandates, legal statues, best practices of national statistical agencies, and past precedent. Harvard University's Privacy Tools Project is working to development an axiomatic framework in this spirit (Wood 2016). Some important axioms for privacy mechanisms have also been discussed briefly in the literature; "Pufferfish" (Kifer and Machanavajjhala 2014) describes a few such axioms (Transformation Invariance, Convexity), for example. A key theme of this discussion is that simple, standard metrics (e.g. *L1/L2* error) are not likely to mirror real-world utility very well, and may generate unwanted artifacts[9] in the simulated microdata.

When using the Multiplicative Weights Exponential Mechanism (MWEM) (Dwork and Roth 2014; Hardt, Ligett, and McSherry 2012) to answer queries, memory is a big hurdle when the universe of record types is high dimensional. The idea is to first create synthetic Decennial Census data via MWEM to get accuracy without an embarrassingly high ε, and then iteratively fit to publicly known statistics in order to improve the initial synthetic microdata. Dan Kifer's group at the Census Bureau is currently working on this approach. The current plan is to start simulating data at the national level, fix things that look bad, then move to the state level (focusing on states whose data differ starkly from the national data), and so on. Starting low and propagating up is ill advised due to resulting error expansion.

There was a suggestion to look at the mostly abandoned literature in controlled rounding as a post-processing tool to achieve consistency after generating tables. Such a technique might be used for ensuring consistency with publicly known statistics or to respect the Census Bureau's large collection of structural zeroes/"edits" (e.g. a daughter may not be older than her father). As an alternative to controlled rounding, it was suggested that a mixture model technique can achieve a low-dimensional representation of high-dimensional data and will place zero probability on structural zero queries, but this approach tends to badly over-estimate some multivariate statistics. Furthermore, although structural zeroes might appear to be useful for reducing the data universe size (one of our other major problems), discussants noted that reducing the size of the universe by removing all structural zeroes is a coding nightmare and unlikely to be sufficient to solve the memory problem.

Although we lack detailed, high-quality usage statistics, at a high level there is a clear priority ordering of table accuracy requirements: PL94 must be the most accurate (with exact counts for the previously indicated quantities), the remaining tables of SF1 are of secondary priority, and the large number of remaining SF2 tables are of tertiary importance. It is worthwhile to note we must accept

[9] More strongly, it was suggested that simulated microdata may contain unusual artifacts (e.g. many regions of implausible uniformity) generally, and that these may be problematic for social scientists investigating PUMS generated with differentially private mechanisms.

whatever privacy loss is incurred through the publication of PL94 population counts due to the Congressional mandate that these be exact.

While considering how to motivate differentially private mechanisms for policy makers, there was a short discussion about what impact an explanation of privacy measures might have on response rates and response accuracy. The key idea here was that, if respondents knew that more stringent privacy methods were in use than for prior censuses, response rates and accuracy might increase as a result, and this improvement might help ease some of the likely policy costs of a formally private approach (e.g. increased noise variance due to disclosure limitation techniques). However, implementing a field test of this idea seems unlikely, so the topic was dropped from discussion.

Returning to our earlier note that debugging is likely to be quite difficult, it was noted that it is difficult to write automated tests for verifying that a given algorithm correctly guarantees some given level of differential privacy (or differential privacy at any level). We should be able to test for some common mistakes (e.g. an unintentionally inverted Laplace parameter) and well-known side channel attacks (e.g. privacy loss due to truncation), but general side-channel attacks are difficult to anticipate and address. Both bugs and side-channel attacks can erode privacy guarantees, so broader verification of the code is important. A code-vetting/ bug-catching/ algorithm critique and/or development contest (run independent of the Census Bureau, e.g. by a university, due to legal concerns) might help to engage the broader community of privacy experts in addressing our major challenges. The 2020 Census timeline must be kept in mind in considering when such a contest might be held. Our central goal is to have code ready for end-to-end test of census disclosure by 2018. The R&M Directorate could then push for an official release of code for vetting prior to 2020.

## References


Dwork, Cynthia, and Aaron Roth. 2014. “The Algorithmic Foundations of Differential Privacy.” *Foundations and Trends in Theoretical Computer Science*. now publishers, Inc., 211–407.

Hardt, Moritz, Katrina Ligett, and Frank McSherry. 2012. “A Simple and Practical Algorithm for Differentially Private Data Release.” In *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 2339–47. Curran Associates, Inc.

Kifer, Daniel, and Ashwin Machanavajjhala. 2014. “Pufferfish: A Framework for Mathematical Privacy Definitions.” *ACM Trans. Database Syst.* 39 (1). New York, NY, USA: ACM: 3:1–3:36.

U.S. Census Bureau. 2012. “2010 Census Summary File 1 - Technical Documentation.” SF1/10-4 (RV). U.S. Census Bureau. http://www.census.gov/prod/cen2010/doc/sf1.pdf.

Wood, Alexandra. 2016. “A Modern Approach to Designing Privacy-Aware Data Releases.” presented at the 13th Biennial Federal Committee on Statistical Methodology (FCSM) Policy Conference, Washington DC, December 7.




Funding through NSF CNS-1012593 and Sloan Foundation.

# 2017 Economic Census

*Based on notes prepared by Scot Dahl, Hang Kim, Jenny Thompson, and Lars Vilhuber*

## Summary of Data Product

The Economic Census is a census of businesses conducted every five years by the U.S. Census Bureau. Forms are sent out to nearly 4 million businesses, broadly representative of the complete U.S. geography and all industries, though some industries are excluded. Respondents are asked to operational and performance data,[10] and their response is required by law.

The economic census is primarily conducted on an establishment basis, where an establishment is defined for a location and activity. Companies are requested 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 a mixture of a complete enumeration for certain types of businesses, and sampling of other types.

The EC collects information from sampled establishments on the revenue obtained from product sales (hereafter referred to as "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 the general statistics variables, but auxiliary data are not available for the other items.

In the 2017 EC, missing product data will be imputed using hot deck imputation (Thompson and Liu 2015; Knutson and Martin 2015), and variance estimates for product totals will be published for the first time. Depending on the industry, random hot deck or nearest neighbor hot deck imputation will be implemented (Tolliver and Betchel 2015; Bechtel, Morris, and Thompson 2015). The variance estimator accounts for sampling variance, calibration weighting, and imputation variance. The imputation variance component is the most challenging to estimate, as composite imputation is employed for general statistics items (Shao and Steel 1999) and the imputation rates for several products is quite high. In addition, the highly stratified sample design makes it difficult to accurately estimate the sampling variance component.

## Summary of Discussion

The key challenge are the changes happening in the 2017 EC (Economic Census):

[10] For examples, see sample forms for MC-31303 Broadwoven Fabrics (https://bhs.econ.census.gov/ec12/mc-31-33/mc-31303-form.html), MI-21101 Crude Petroleum and Natural Gas Extraction (https://bhs.econ.census.gov/ec12/mi-21/mi-21101-form.html), and RT-44101 Automobile Dealers (https://bhs.econ.census.gov/ec12/rt-44-45/rt-44101-form.html).

1. All electronic data collection

2. Use of North American Product Classification System (NAPCS)

3. New sample designs for some sectors (Manufacturing, Mining, and Construction)

There was an initial discussion surrounding NAPCS and the American Industry Classification System (NAICS). The NAPCS classification procedure is a major departure from the current collection procedures which explicitly links product codes to industry. NAPCS will be introduced in the 2017 Economic Census (EC), and economy-wide product tabulations from cross-sector publications will be released for the first time. The group also briefly discussed EC data collection challenges. Some variables are obtained for all establishments (general statistics variables: receipts, payroll, employment, etc.), either through direct collection or administrative data substitution. Other data are collected only from sampled establishments.

The key disclosure limitation challenge that the team will focus on is the disclosure limitation process for NAPCS (product) estimates, given the 2017 EC editing and imputation procedures. 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.

## Synthetic data generation project for 2017 EC

Beginning in 2017, the Center for Disclosure Avoidance Research (CDAR) is commissioning an interdisciplinary team led by Hang Kim (University of Cincinnati) 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 that satisfies edit rules and disclosure constraints on historic EC data from the manufacturing sector. The team hopes to extend their multivariate normal joint model to accommodate these data sets, although other modifications or models may need to be investigated. For example, EC synthetic data has an additional constraint, 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 estimates.

Besides 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 think about actual usage and analysis by outside users. How is confidence in the synthetic data created? One possibility is a verification server (Reiter, Oganian, and Karr 2009).

The group also discussed the relative merits of fully synthetic vs. partially synthetic data. The advantages of fully synthetic data – consistent properties for all estimates, preferred by some users - are

weighed against the disadvantages – some outliers are synthesized too well and may pose a disclosure risk.

## Miscellaneous

Quantification of data quality is generally difficult: it remains an open question whether one should rely on a priori, formal privacy guarantees (such as differential privacy), or ex post risk measures often used with a synthetic data approach (Kinney et al. 2011).

## References


Bechtel, Laura T., Darcy Steeg Morris, and Katherine Jenny Thompson. 2015. "Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study." http://sites.usa.gov/fcsm/files/2016/03/E1_Bechtel_2015FCSM.pdf.

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

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." *International Statistical Review = Revue Internationale de Statistique* 79 (3). Blackwell Publishing Ltd: 362–84.

Knutson, Jeremy, and Jared Martin. 2015. "Evaluation of Alternative Imputation Methods for Economic Census Products: The Cook-Off." In *Proceedings of the Section on Survey Research Methods, American Statistical Association*. https://ww2.amstat.org/meetings/JSM/2015/onlineprogram/AbstractDetails.cfm?abstractid=315426.

Reiter, Jerome P., Anna Oganian, and Alan F. Karr. 2009. "Verification Servers: Enabling Analysts to Assess the Quality of Inferences from Public Use Data." *Computational Statistics & Data Analysis* 53 (4): 1475–82.

Shao, Jun, and Philip Steel. 1999. "Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions." *Journal of the American Statistical Association* 94 (445). [American Statistical Association, Taylor & Francis, Ltd.]: 254–65.

Thompson, Katherine Jennifer, and Xijian Liu. 2015. "On Recommending a Single Imputation Method for Economic Census Products." In *Proceedings of the Section on Government Statistics, American Statistical Association*. https://ww2.amstat.org/meetings/jsm/2015/onlineprogram/AbstractDetails.cfm?abstractid=315412.

Tolliver, Kevin, and Laura Betchel. 2015. "Implementation of Hot Deck Imputation on Economic Census Products." In *Proceedings of the Section on Survey Research Methods, American Statistical Association*. https://ww2.amstat.org/meetings/jsm/2015/onlineprogram/AbstractDetails.cfm?abstractid=315511.




Funding through NSF CNS-1012593 and Sloan Foundation.

# Appendix

## Agenda for the October 14, 2016 workshop

| *Start* | *Duration* | *Topic* | *Location* |
|---|---|---|---|
| 8:30 | (0h30) | **Welcome, housekeeping plan, Associate Director's remarks** | 8h112 |
| 9:00 | (2h30) | **Break out working groups** | |
| | | *(I) American Community Survey* | 8h112 |
| | | *(II) LODES* | B |
| 11:30 | (0h30) | **Plenum - Interim summary** | 8h112 |
| 12:00 | (1h30) | *Lunch* | 8h112 |
| 13:30 | (2h30) | **Break out working groups** | |
| | | *(III) Census 2020* | 8h112 |
| | | *(IV) Economic Census* | B |
| 16:00 | (1h00) | **Plenum - Summary** | 8h112 |
| 17:00 | | End of workshop | |

# Participants

| *Participant* | *Affiliation* | *Website* |
|---|---|---|
| Lars Vilhuber | Labor Dynamics Institute, Cornell University | https://www.vilhuber.com/lars/ |
| John M. Abowd | U.S. Census Bureau | https://courses.cit.cornell.edu/jma7/ |
| Ian M. Schmutte | University of Georgia | http://people.terry.uga.edu/schmutte/ |
| Nissim Kobliner, Yaacov | Harvard University | http://crcs.seas.harvard.edu/people/kobbi-nissim |
| Gerome Miklau | University of Massachusetts - Amherst | https://people.cs.umass.edu/~miklau/ |
| Jerry Reiter | Duke University | http://www2.stat.duke.edu/~jerry/ |
| Ashwin Machanavajjhala | Duke University | https://users.cs.duke.edu/~ashwin/ |
| Daniel Kifer | Penn State University | http://www.cse.psu.edu/~duk17/ |
| William Sexton | Labor Dynamics Institute, Cornell University | https://www.linkedin.com/in/william-sexton-239b7937 |
| Vishesh Karwa | Penn State University and Harvard University | http://www.personal.psu.edu/vkk106/ |
| Hang Kim | University of Cincinnati | http://www.artsci.uc.edu/departments/math/fac_staff.html?eid=kim3h4&thecomp=uceprof |
| Michael Hay | Colgate University | http://www.colgate.edu/facultysearch/FacultyDirectory/michael-hay |
| Chris Clifton | Purdue University | https://www.cs.purdue.edu/people/clifton |
| Rolando Rodriguez | U.S. Census Bureau | |
| Amy Lauger | U.S. Census Bureau | |
| Phil Leclerc | U.S. Census Bureau | |
| Aref Dajani | U.S. Census Bureau | |
| Jenny Thompson | U.S. Census Bureau | |
| Matthew Graham | U.S. Census Bureau | |
| Mark Kutzbach | U.S. Census Bureau | |
| Scot Dahl | U.S. Census Bureau | |

Seated: Chris Clifton, William Sexton, Matthew Graham, Ashwin Machanavajjhala, Dan Kifer, John Abowd. Standing: Jenny Thompson, Phil Leclerc, Vishesh Karwa, Hang Kim, Amy Lauger, Rolando Rodriguez, Scot Dahl, Jerry Reiter, Aref Dajani, Michael Hay, Lars Vilhuber, Ian Schmutte, Gerome Miklau. Not pictured: Mark Kutzbach, Kobbi Nissim


Funding through NSF CNS-1012593 and Sloan Foundation.