9-20-2017

Proceedings from the 2017 Cornell-Census- 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 2017 Cornell-Census- NSF-Sloan Workshop on Practical Privacy

Abstract
These proceedings report on a workshop hosted at the U.S. Census Bureau on May 8, 2017. Our purpose was to gather experts from various backgrounds together to continue discussing the development of formal privacy systems for Census Bureau data products. This workshop was a successor to a previous workshop held in October 2016 (Vilhuber & Schmutte 2017). At our prior workshop, we hosted computer scientists, survey statisticians, and economists, all of whom were experts in data privacy. At that time we discussed the practical implementation of cutting-edge methods for publishing data with formal, provable privacy guarantees, with a focus on applications to Census Bureau data products. The teams developing those applications were just starting out when our first workshop took place, and we spent our time brainstorming solutions to the various problems researchers were encountering, or anticipated encountering. For these cutting-edge formal privacy models, there had been very little effort in the academic literature to apply those methods in real-world settings with large, messy data. We therefore brought together an expanded group of specialists from academia and government who could shed light on technical challenges, subject matter challenges and address how data users might react to changes in data availability and publishing standards.

In May 2017, we organized a follow-up workshop, which these proceedings report on. We reviewed progress made in four different areas. The four topics discussed as part of the workshop were 1. the 2020 Decennial Census; 2. the American Community Survey (ACS); 3. the 2017 Economic Census; 4. measuring the demand for privacy and for data quality.

As in our earlier workshop, our goals were 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.

Comments can be provided at https://goo.gl/ZAh3YE





Proceedings from the 2017 Cornell-Census-NSF-Sloan Workshop on Practical Privacy

Held on Monday, May 8, 2017 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.

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. All note takers were academics, and not Census Bureau employees, and the notes have been summarized by the editors. All results have been reviewed to ensure that no confidential information is disclosed.

Acknowledgement

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.

Goals and Methods of the Workshop

These proceedings report on a workshop hosted at the U.S. Census Bureau on May 8, 2017. Our purpose was to gather experts from various backgrounds together to continue discussing the development of formal privacy systems for Census Bureau data products. This workshop was a successor to a previous workshop held in October 2016 (Vilhuber & Schmutte 2017). At our prior workshop, we hosted computer scientists, survey statisticians, and economists, all of whom were experts in data privacy. At that time we discussed the practical implementation of cutting-edge methods for publishing data with formal, provable privacy guarantees, with a focus on applications to Census Bureau data products. The teams developing those applications were just starting out when our first workshop took place, and we spent our time brainstorming solutions to the various problems researchers were encountering, or anticipated encountering. For these cutting-edge formal privacy models, there had been very little effort in the academic literature to apply those methods in real-world settings with large, messy data. We therefore brought together an expanded group of specialists from academia and government who could shed light on technical challenges, subject matter challenges and address how data users might react to changes in data availability and publishing standards.

In May 2017, we organized a follow-up workshop, which these proceedings report on. We reviewed progress made in four different areas. The four topics discussed as part of the workshop were

1. the 2020 Decennial Census;
2. the American Community Survey (ACS);
3. the 2017 Economic Census;
4. measuring the demand for privacy and for data quality.

As in our earlier workshop, our goals were 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 met as a group in four sequential sessions. In each session, one research team presented on its approach to data modeling, their progress to date, and any challenges they were facing in developing practical implementations. Every session was assigned two notetakers who recorded the discussion according to the Chatham House rule.[1] The entire group was free to discuss any aspect of theory, implementation, etc. No conclusion needed to be reached. The note-takers subsequently drafted summaries of the discussions, which were circulated among the group members for review and correction. The final summary appears in these proceedings.

Common Themes

Several themes recurred throughout the workshop. All three data products under discussion involve some kind of hierarchical structure. For example, in the ACS and Decennial Census, individuals are nested within households, and it is important to keep variables consistent across individuals in the same household. Also, all data products give rise to "structural zeros" -- that is, combinations of variables that can never be jointly observed because of logical constraints. Both hierarchical structures and structural zeros are hard to incorporate into synthetic data models; particularly those with formal privacy guarantees.

The teams working on ACS and Decennial Census both described serious computational challenges. While some of these computational challenges are fundamental, others can be overcome, or at least alleviated, with the correct technology. There seems to be a need to speed the ability of the Census Bureau to acquire and install up-to-date software and computing infrastructure for research and development.

Finally, there is a tension between the data models used for synthesis and formal privacy ,and data processing and editing. All the teams noted that it was important to account for weighting, missing data imputation, and post-processing edits. However, the teams have deferred consideration of these complications to the future. This is certainly a reasonable

[1] https://www.chathamhouse.org/about/chatham-house-rule

decision given the current state-of-the-art. However, it suggests very clear directions for theoretical and applied research.

Next Steps

The participants found the workshop helpful and many expressed an interest in meeting again in the near future. The group also introduced the possibility of developing a network that could facilitate ongoing discussion and collaboration across the teams.

Session on 2020 Decennial Census

Overview

The team tasked with implementing formal privacy for the 2020 Decennial Census discussed its planned approach at the Fall 2016 Workshop on Practical Privacy (Vilhuber & Schmutte 2017). What follows is a brief overview of their charge. The Census Bureau is conducting an overhaul of disclosure avoidance methods used to protect data publications based on the 2020 Decennial Census. A specific goal of this overhaul is to deliver data to the public with a formal privacy guarantee. To achieve this goal, the team is attempting to generate synthetic microdata that are differentially private. The logic of this approach is that the differentially private synthetic microdata, and publication tables built from it, will all have the same formal privacy guarantee.

Several additional objectives constrain what this team needs to deliver. First, the microdata should appear familiar to both internal and external stakeholders. Second, they need to provide a compact representation of query answers. Finally, and crucially, the data must deliver key sets of publication tables ("PL94" and "SF1" tables, discussed below) that are mutually consistent in the sense that they satisfy "adding up" constraints along the table margins. Once the team has developed a technology for producing formally private microdata, it will be up to policy makers to choose the level of privacy to provide, knowing that increasing privacy necessarily entails a measurable loss in the accuracy of the published data.

Tables produced to satisfy Public Law 94-171, the *PL94 tables* (Anon 2011), contain counts of the total and voting age population by race and ethnicity, along with counts of housing units. These counts are published at several different nested and non-nested levels of geography, including at the level of the Census block. Because the Census Bureau is required by federal statute to provide these tables to state governments for congressional redistricting, they are an extremely high priority product.

As part of the PL94 tables, there are 3 independent counts that are released exactly as enumerated at the block level: the total count of householders, the total voting age population,

and the total population. The justice department ruled that it is acceptable to release unperturbed voting age counts at the block level (Anon 2011). In 2000 and 2010 the unoccupied household counts are also exact. This matter has not been settled for 2020 yet. An overarching goal with respect to stakeholders is to provide them with microdata that maintains the fidelity of key PL94 tables. Specifically, the exact block-level counts described above must be preserved. Maintaining privacy while publishing these exact counts is difficult, since they impose restrictions that can be exploited in post-processing.

Summary File 1 (SF1) reports detailed summaries of all questions asked in the Decennial Census. There are a very large number of such tables. A bulk of development time has been spent working on the SF1 tables. Given the balance between detail and priority, these tables are currently the highest focus. Importantly, SF1 contains both individual and household tables.[2]

Current Approach

The group has attempted to produce synthetic microdata using differentially private mechanisms. They have worked with basic approaches such as Laplace, Geometric, and Exponential mechanisms. There has been some difficulty accessing the `crlibm` library (de Dinechin et al. 2008) on Census Bureau servers, which is required for proper sampling from the Laplace distribution. More complicated mechanisms, PrivTree, MWEM, and NoiseDown, have also been tested and subsequently dismissed. None of these algorithms compete with the error levels produced by HB Tree, which is the algorithm currently in active use. The Matrix Mechanism is also still under consideration. As in the ACS data project discussed in the next section, the current approach is to generate synthetic data assuming data edits will occur as a type of post-processing.

Structural zeros are common in the data schema, and these can be very difficult to incorporate in modeling. A visualization team produces and reviews depictions of the synthetic data that marginalize across one variable, which allows them to check whether structural zeros are controlled properly. In general, the structural zeros should be incorporated in the process

[2] Summary File 2 contains the greatest level of geographic detail but is seen as a low priority.

generating the differentially private synthetic data, rather than being enforced through post-processing.

Some consideration has been given to establishing a backend validation server to allow users to get some information about how far the published data are from the confidential data. However, releasing actual accuracy expenses some of the privacy budget. Historically only undercount, overcount are reported. It may be worth the privacy cost to build confidence among users in the quality of the published data.

Several postprocessing/inference methods have also been considered. These include ordinary least squares, nonnegative least squares, and mixed integer programs. Ordinary least squares provides nice closed forms for mean-squared error, but can't guarantee nonnegativity or integer counts. Nonnegative least squares (NNLS) solves the nonnegativity issue but does not give integer counts. Additionally, small biases in individual cells compound for aggregates, though targeting all queries over all ranges seems to solve the bias issue. Mixed integer linear programming allows for both non-negativity and can yield integer counts but without a nice closed form solution. This method is CPU and memory-intensive.

Major Research Challenges

The workload consists of publishing the many thousands of tables in SF1 and PL94. These are to be partitioned into groups. Important subsets include individual tables, household tables, and group quarter. The population in households of types is highly sensitive as it is especially responsive to individual's changing type. Currently, there are no plans to handle nonlinear queries within the privacy algorithms but rather relegate this task to postprocessing.

Group quarters pose a particular problem. For example, prison counts need to match exactly. For this reason, group quarters have historically not been swapped and protection was only afforded to aggregate types. However, it is unclear how to best handle group quarters within the framework of differential privacy.

So far, the differential privacy models under consideration have focused on data at the individual level. The question of how to define and manage privacy for household records

remains open. It was surmised that correlations within household compromise individual-level privacy.

In general, future research will need to solve the problem of how to allocate synthetic individuals into synthetic households. This is followed by the problem of how to allocate synthetic households to states.

Additional Discussion

The Center for Disclosure Avoidance Research (CDAR) within the Census Bureau must assess whether particular queries can be handled by differential privacy. For example, in some cases a more detailed workload consisting of all range queries may result in greater accuracy than just using the SF1 range queries. Identifying the correct sparsity structure may make it easier to answer queries but an efficient way to spending the privacy budget to discover this structure remains an open question.

There was a further wide-ranging discussion of the issues elaborated above. One suggestion was to allocate the privacy budget to ensure error in national tables is reasonable. The national tables are widely checked, and good performance there seems necessary to build confidence in the user community.

There was also a brief discussion of whether noise added from editing and other post-processing steps, which contribute to total survey error, can be incorporated into formal privacy measures. Defining this is an open question for theoretical research.

Some of the differential privacy algorithms do not have a well-defined bound on the errors in data quality they introduce. Furthermore, the L1 and L2 loss functions common in the literature are not universally appropriate ways to measure data quality.

Session on American Community Survey (ACS)

Overview

In this session, participants discussed the ongoing effort to implement novel privacy-preserving statistical methods for the American Community Survey (ACS). Please see the original discussion of this project in the Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy (Vilhuber & Schmutte 2017). The discussion centered around progress-to-date and open challenges.

Current efforts are focused on developing a methodology for formally private synthesis of microdata. The synthesized microdata can be used to generate the ACS data products, which include summary tables and public use microdata samples (PUMS). It is important that the implementation fit into the current ACS production timeline and process and that it meets the needs of ACS stakeholders. Thus, this project requires close collaboration with ACS staff at the Census Bureau.

The team has made progress on a method for producing synthetic ACS microdata. This methodology was reviewed in great detail during the workshop. Modifying the approach to include formal privacy protection is still a work in progress.

Before getting into the details of data synthesis approach, participants reviewed those features of ACS that make developing synthetic data particularly challenging:

- Large number of variables. There are 200 variables to synthesize that include Census variables, variables related to housing characteristics (type of housing, when built, worth, etc.) and additional person variables (education level, etc.).
- Skip patterns create structural zeros. A skip pattern occurs when the answer to one question renders other questions inapplicable (e.g., a

householder may be asked different questions depending on whether he/she owns a home). Skip patterns complicate data synthesis because certain combinations of responses are impossible (a form of structural zero).

- There is a high rate of survey non-response (34% = 1.2M out of 3.5M households per year). In current practice, the Bureau adjusts survey weights to adjust for non-response. The data synthesis should address weighting somehow.
- Geography and sample size. There is a desire to retain accurate estimates at multiple granularities of geography. At the tract level (finest granularity), the sample sizes are small in any given year. For this reason tract-level estimates are aggregates over 5 years.

Current Approach

The participants then reviewed the progress to date. The current approach to data synthesis is based on a model described in (Hu et al. 2017). This approach addresses two key challenges of synthesizing the ACS:

- Categorical attributes are challenging to model well. Identifying the appropriate set of interactions and handling sampling zeros compounds this difficulty.
- Individuals are organized into households. But there are within-household relationships that constrain the feasible set of values (e.g., constraints on ages implied by parent-child relationships). The nested structure of categorical variables has been a focus of the current modeling effort.

The synthesis model has two main steps. First, synthetic housing data is created using chained regressions (regression on variable i given sampled values for variables 1, ..., i-1). Then the houses are populated with synthetic individuals. The model has two levels of latent classes. Households belong to latent household classes (≈ 15 total) and household members belong to individual classes (≈ 10 per household class). The individual model is

conditioned on membership in the corresponding household class. Variables are sampled independently, given the latent class membership of the household and individual.

Given this conditional independence structure, the model places support on household configurations that are practically speaking impossible (e.g., a parent being younger than a child). To ensure that the synthetic data only includes feasible configurations, the model is fit to the data using a variant of MCMC that rejects infeasible samples. This is computationally expensive because the model assigns a large probability to the infeasible space. Hence, a very large number of samples are required to get a sufficient quantity that are not rejected by the feasibility constraint.

The current approach has been evaluated using 2012 PUMS data for 10,000 households. Only one year was included because of computational issues associated with sampling larger datasets. Evaluation is based on comparing confidence intervals for statistics of interest between the synthetic and the original data. Overall, the confidence intervals are quite close for many household-level statistics but become less accurate for statistics that characterize household composition (e.g., proportion of three-generation family households) as well as relationships between individuals within a household (e.g., age difference between spouses). In addition, it appears as though statistics related to home ownership (specifically "White couple own" and "Non-white couple own") were not fit well. A possible explanation is that these statistics require capturing a relationship between household variables (ownership) and individuals within the home (e.g. their races).

Major Research Challenges

With the latent class model just described, there remain many outstanding challenges. First, there are several features of the original data that are not addressed with the current approach. Several are particularly important, and will be a focus of research effort moving forward:

- Geography. The current approach generates data at the state level. To obtain finer granularity in geographic detail, the plan is to build a model of position (lat/long) given other variables.
- Editing and imputation of missing values. The plan is to incorporate imputation into the modeling but maintain editing as a post-processing step.
- How to incorporate sampling weights has not been addressed yet.

Additional Discussion

The group discussed many of the computational and modeling issues raised by the work-to-date. First, the problem of synthesizing the full ACS requires being able to scale this approach. However, the rejection sampling approach is very time-consuming. While sampling can be sped up through parallelization, a question was raised about whether the linear speedup that parallelization could offer is capable of combatting what may be exponential complexity. The complexity increases with household size so one suggestion was to break up computation based on household size and possibly design different strategies for larger households.

The group wondered about the feasibility of embedding more restrictions into the model itself. This is a possible direction but requires specifying valid data configurations in advance. Previous effort of this type by members of the research team suggests that valid configurations are complicated to specify in advance and difficult to incorporate into the model.

It was also observed that confidence intervals around variables with very small frequency of occurrence, measuring small population proportions (e.g., three-generation-family), were biased upwards. There was a conjecture that such error arises from poor fit to extreme values, which may be desirable from a privacy perspective.

There was some discussion about the mixture model approach and the ability to modify the model in response to problems with fit. In brief, because the latent mixture components are interdependent, if the model fits poorly in some dimension, it's not always clear how to adjust the model to improve the fit. A

change made to address one issue can hurt the fit elsewhere. This is a challenge with this approach.

Structural zeros were discussed at length. Specifically, how are the structural zeros identified and enforced? It seems structural zeros were established both by analyzing the variables (and using common sense) and by analysis of patterns in the data. It is possible, however, that some variable combinations that should be structural zeros are actually non-zero in the synthetic data. This is an area for further refinement, both of the model, and of the tools for determining where structural zeros should occur.

Sampling zeros were also discussed. The model places support on sampling zeros but it's unclear whether the probability of their occurrence is large enough. Some said that the probability mass was less important but that sparsity patterns in the synthetic data generally match the sparsity of real data. This is an area that needs further investigation. The challenges of identifying sparsity patterns under differential privacy were briefly discussed.

Due to time constraints, the session ended before the efforts on disclosure risk and differentially private synthesis could be discussed.

Session on 2017 Economic Census

Overview

In this session, participants discussed ongoing work and a plan to generate synthetic data for the 2017 Economic Census. This activity is in its initial phases.

The discussion began with a thorough overview of the Economic Census. Part of this summary was included in the previous workshop proceeding (Vilhuber & Schmutte 2017). Briefly, the Economic Census is moving to all electronic data collection with the 2017 survey. It is also moving from the NAICS industrial classification system to a product-based system (NAPCS) that classifies all products the establishment sells. As has always been the case, the Economic Census samples some establishments with certainty while other, smaller, units are sampled probabilistically. There are four core general statistics to be collected from all units. Finally, the Economic Census will continue to use a hot deck imputation to generate post edit-imputation data. The team's purpose is to develop confidentiality protection and generate synthetic data based on the post edit-imputation data.

Current Approach

A nonparametric Bayesian model (Kim et al. 2016) serves as the framework to generate synthetic microdata. As in the ACS application, the current approach will produce synthetic data that do not have a formal privacy guarantee. Augmenting this approach to render the synthetic data differentially private is a topic for future work.

The synthetic data have a calibration requirement. Specifically, the microdata must preserve, up to some error, totals that appear in publication tables based on the confidential data. This calibration requirement will allow some inconsistencies, i.e., the calculated sum does not need to be exactly the

same with the total. The current approach is to use penalty functions (interpreted as prior distributions as in Bayesian lasso), which will be incorporated in the synthetic data model. Finally the current approach does allow for modeling of part-year reporters.

Major Research Challenges and Discussion

About published margin (or known total) in calibration, there was a question about “what is or where exists the known total in practice”. There was also some discussion of the hot deck imputation procedure in which it was remarked that the hot deck has some quality issues. However, the hot deck will be maintained for the next round of the Economic Census, so the synthetic data must respect this feature of the data generating process.

There was some discussion as well about cell suppression and whether geography is a factor in such suppression.

With respect to the data model, there was a long discussion about the nature of economic data. Specifically, the joint distribution of economic variables often does not follow a simple parametric distribution. In addition, there exist complicated edit rules. Finally there are few strong predictors for many variables. These pose challenges for editing and imputation processes but also in building a synthesis model that preserves joint distribution and satisfies edit rules. For example, in collection rules, the system will allow users to put incorrect or missing values for total values (receipts). Also, across all service industries, a few number of major products account for high percentage of total revenue.

Session on Demand for Privacy

Overview

In this session, participants discussed how to provide guidance for policymakers in determining the level of privacy protection to provide. As background, the group discussed a recent working paper (Abowd & Schmutte 2017) that builds an economic framework for determining the optimal level of privacy protection and data quality. They show that this problem requires determining how much a policy maker (or "social planner") should be willing to increase accuracy of published data, when doing so requires individuals to sacrifice in terms of foregone, or lost, privacy. Using their model as a starting point, the group discussed different strategies for measuring the value of data quality and the cost of privacy that could be lost when data are published using a formally private mechanism.

Abowd and Schmutte use attitude measures from several different surveys as proxies for underlying preferences for privacy and data accuracy. The measures come from the Cornell National Social Survey (Cornell University 2014) and the Federal Statistical System Public Opinion Survey (Childs et al. 2015) conducted by the Census Bureau's Center for Survey Methodology (CSM).

Current Approach

Research is currently underway in collaboration with CSM to obtain better measures of preferences for privacy and accuracy that can be used to guide policy. The discussion at this workshop serves, in part, as a brainstorming session to help shape this research. The team has planned to initiate a set of opinion surveys that get more directly at individual preferences. First, they will solicit opinions that reflect demand for privacy of the type that can be offered by formal privacy systems. Second, they will solicit opinions that reflect the demand for data accuracy, or the willingness to pay for population statistics of a particular level of quality.

The concepts involved in this study have traits of “public goods”. For example, Census Bureau statistics can be used by anyone without affecting their availability and quality for other users. Hence, measuring individual preferences for such goods is complicated by the fact that such goods appear to the individual to be freely available. Hence, the group is drawing on approaches to measuring willingness to pay for public goods from the economics and social psychology literatures. Generically in models with public goods it is very hard to quantify benefits of things with very large externalities, like road networks or environmental protection.

Furthermore, because the concepts involved are complex and unfamiliar to most citizens, the team plans to do qualitative testing ahead of any broader survey to ensure that questions are posed in such a way that they measure, as closely as possible, the theoretical concepts of interest. At the time of the workshop, the plan was to roll out qualitative interview-based testing along with a pilot study using a convenience sample from Amazon Mechanical Turk, Google Consumer Surveys, or similar online survey engines.

Discussion

The discussion was thorough and wide-ranging. Several members of the group referred to Acquisti’s work on the analysis of willingness to pay/willingness to accept in the context of online privacy (e.g. Acquisti et al. 2013). Relatedly, there is ongoing work, funded by DARPA, between UMass and Carnegie Mellon on communicating to laypeople about differential privacy. It was noted, however, that Acquisti’s work is more about the whether or not people choose to disclose information, and in this context people usually think of identity theft. This may be different than measuring people’s attitudes toward risk of privacy loss after disclosure.

There was, in general, a deep concern in the group about how to properly articulate privacy and data quality concepts to the general public. Specifically, the working group gathered for this workshop is used to a particular mathematical definition of privacy. However, we have very little sense of what the public thinks of that definition and the extent to which it corresponds to how people think about privacy in general. These are complicated concepts, so any survey research needs to make clear whether people understand them. Furthermore, it is important to be careful with language, because framing questions, say in terms of identity theft, can lead to very different responses than if privacy is discussed in some other context. The team might

consider whether the problem could be more usefully framed in terms of “inferential privacy” rather than “differential privacy.”

One way to deal with these issues would be to run different studies on the general population and on expert users of data. Different groups will have different ability to comprehend, but also different preferences, and should be studied in isolation. Also, the emphasis on individuals as respondents may be misplaced. A firm-level survey could get into a context where the privacy and data quality concerns are both more salient.

The suitability of Amazon Mechanical Turk for pilot testing was called into question. Ming Yin at Harvard[3] has some work showing Mechanical Turk surveys can be unreliable in certain settings. Other tools are available that may be more appropriate.

It was observed that there is a seeming asymmetry between

- the demand for privacy, which comes from the individual, and
- the demand for data accuracy, which comes from the analyst.

However, rather than pointing to a flaw in how this work is conceived, this observation may instead highlight a need to clarify where the demand for accuracy arises. That is, what social need are existing uses of public data satisfying? What decisions are they simplifying and why? A very similar discussion arose around a distinction/asymmetry between

- the pain of giving up information which may be immediate and direct, and
- the benefit from higher-quality data which is more abstract / indirect.

Again, this distinction is not as clear as first appears. In some contexts, the pain of disclosure might be remote, indirect, and very probabilistic, and the benefits could potentially be more immediate.

As an alternative approach to measuring preferences, one way to offer policy recommendations would be to start from the option in which there is no (additional) formal privacy, which corresponds to publishing the maximum possible data quality. It may turn out that the total survey error introduced by sampling, editing, etc. provides some ambient privacy protection. Whether this provides a formal privacy guarantee and how much is a theoretical

[3] http://people.seas.harvard.edu/~myin/

question. Nevertheless, with this in mind, the key question is not necessarily “what is the optimal level of privacy protection”, but “is the optimal level of privacy protection non-zero?” The latter should be simpler to answer than the former.

Participants

Participant	*Affiliation*	*Website*
Aleksandra Slavkovic	Penn State University	http://stat.psu.edu/people/abs12
Aref Dajani	U.S. Census Bureau	
Ashwin Machanavajjhala	Duke University	https://users.cs.duke.edu/~ashwin/
Chris Clifton	Purdue University	https://www.cs.purdue.edu/people/clifton
Daniel Kifer	Penn State University and U.S. Census Bureau	http://www.cse.psu.edu/~duk17/
Gerome Miklau	UMass Amherst and U.S. Census Bureau	https://people.cs.umass.edu/~miklau/
Hang Kim	University of Cincinnati	https://goo.gl/eX7wpv
Ian M. Schmutte	University of Georgia	http://people.terry.uga.edu/schmutte/
Jennifer Childs	U.S. Census Bureau	
Jenny Thompson	U.S. Census Bureau	
Jerry Reiter	Duke University and U.S. Census Bureau	http://www2.stat.duke.edu/~jerry/
Joerg Drechsler	Institute for Employment Research, Germany	http://iab.de/123/section.aspx/Mitarbeiter/505
John M. Abowd	U.S. Census Bureau and Cornell University	https://courses.cit.cornell.edu/jma7/
Joshua Snoke	Penn State University	http://www.joshuasnoke.com/
Kobbi Nissim	Georgetown University	http://crcs.seas.harvard.edu/people/kobbi-nissim
Lars Vilhuber	Cornell University and U.S. Census Bureau	https://www.vilhuber.com/lars/
Matthew Graham	U.S. Census Bureau	
Michael Freiman	U.S. Census Bureau	
Michael Hay	Colgate University and U.S. Census Bureau	http://www.colgate.edu/facultysearch/FacultyDirectory/michael-hay
Phil Leclerc	U.S. Census Bureau	
Rolando Rodriguez	U.S. Census Bureau	
Simson Garfinkel	U.S. Census Bureau	https://simson.net/
Vishesh Karwa	Harvard University	http://www.personal.psu.edu/vkk106/
William Sexton	Cornell University and U.S. Census Bureau	https://www.linkedin.com/in/william-sexton-239b7937

References

Abowd, J.M. & Schmutte, I., 2017. *Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods*, Labor Dynamics Institute, Cornell University. Available at: http://digitalcommons.ilr.cornell.edu/ldi/37.

Acquisti, A., John, L.K. & Loewenstein, G., 2013. What Is Privacy Worth? *The Journal of legal studies*, 42(2), p.1. Available at: http://chicagounbound.uchicago.edu/jls/vol42/iss2/1 [Accessed September 19, 2017].

Anon, 2011. *2010 Census Redistricting Data (Public Law 94-171) Summary File: Technical Documentation*, US Census Bureau. Available at: https://www.census.gov/prod/cen2010/doc/pl94-171.pdf.

Childs, J.H., King, R. & Fobia, A.C., 2015. Confidence in U. S. Federal Statistical Agencies. *Survey Practice*, 8(5). Available at: http://www.surveypractice.org/index.php/SurveyPractice/article/view/314/html_38 [Accessed September 19, 2017].

Cornell University, 2014. Cornell National Social Survey (CNSS) Integrated. Beta version. Available at: http://doi.org/10.3886/E100424V1.

de Dinechin, F., Muller, J.-M. & Revy, G., 2008. CRlibm: a Library of Elementary Functions with Correct Rounding. Available at: http://raweb.inria.fr/rapportsactivite/RA2009/arenaire/uid45.html [Accessed September 19, 2017].

Hu, J., Reiter, J.P. & Wang, Q., 2017. Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data. Available at: http://projecteuclid.org/euclid.ba/1485227030.

Kim, H.J., Reiter, J.P. & Karr, A.F., 2016. Simultaneous edit-imputation and disclosure limitation for business establishment data. *Journal of applied statistics*, 0(0), pp.1–20. Available at: http://dx.doi.org/10.1080/02664763.2016.1267123.

Vilhuber, L. & Schmutte, I.M. eds., 2017. *Proceedings from the 2016 NSF–Sloan Workshop on Practical Privacy*, Labor Dynamics Institute. Available at: http://digitalcommons.ilr.cornell.edu/ldi/33/ [Accessed June 30, 2017].