Scholars and experts from multiple disciplines will comment on the findings from Part (I), particularly focusing on implications for social science research and insights into the unavoidable trade-off between data privacy and data utility.

Chair: Erica Groshen
Cornell University

Cynthia Dwork
Computer Science, Harvard University

Ori Heffetz
Economics, Cornell University and Hebrew University of Jerusalem

V. Joseph Hotz
Economics, Duke University

Salil Vadhan
Computer Science, Harvard University

Ruobin Gong
Statistics, Rutgers University

Mark Hansen
Journalism, Columbia University

Paul Ohm
Law, Georgetown University
Session II Introduction
Panel on Differential Privacy and the 2020 Census

Erica L. Groshen

Harvard Data Science Review Conference
Friday, October 25, 2019
Differential Privacy for 2020 US Census: Session II

Interpret Session I findings
   – Impact on social science research
   – Trade-offs between data privacy and data utility

Agenda
1. 7 Experts Comment
2. Moderator Qs
3. Audience Qs
Expert Panel
(in order of speaking)

1. Cynthia Dwork (Computer Science, Harvard U.)
2. Ori Heffetz (Economics, Cornell U. & The Hebrew U. of Jerusalem)
3. V. Joseph Hotz (Economics, Duke U.)
4. Salil Vadhan (Computer Science, Harvard U.)
5. Ruobin Gong (Statistics, Rutgers U.)
6. Mark Hansen (Journalism & Statistics, Columbia U.)
7. Paul Ohm (Law, Georgetown U.)
Differential Privacy for 2020 US Census (II)

Cynthia Dwork
Harvard University
Privacy v. Accuracy: A False Dichotomy

- Expand the space of mitigators: eg, allocate funds for ages $\leq 19$
- Why *should* privacy be free?
Understanding Magnitudes

“Less than the sampling error”

- Less than the coverage error?
- Less than shifts in population over one year?
Prioritizing Goals, Designing Measurements

- Strengthen basis for enforcement of Voting Rights Act
- Less sensitive, more robust, techniques
Resist Mission Creep

- The estimated citizenship bits will need protection, diverting precious privacy budget, or exceeding that set by the Data Stewardship Executive Policy Committee
Differential Privacy for 2020 US Census (II)

Ori Heffetz
Cornell University and The Hebrew University of Jerusalem
Differential Privacy for 2020 US Census (II)

V. Joseph Hotz
Duke University
Comments on Assessments of Differential Privacy, 1940 Census Data & Beyond

V. Joseph Hotz
Duke University

Presentation at
Harvard Data Science Review Inaugural Symposium
Harvard University
October 25, 2019
My Perspective/Questions

• How can we use these studies and findings to improve application of disclosure avoidance methods (e.g., differential privacy) to 2020 Census public release data?

• What can we do going forward to improve tradeoffs between information accuracy and the public’s privacy?
Some things I/we learned

• Very useful set of studies.

• Confirmed some things we already suspected.
  • Distortion minimal from Noise-Infusion for more-populated geographic areas & less-rare subgroups.
  • Problem is with sparsely-populated areas & rare-subgroups.

• Things I didn’t think of (but should have).
  • Challenges for using these data to develop reliable sampling frames & costs of inaccuracies.

• Possible discrepancies between non-DP data, e.g. PUMS and NHGIS (National Historic Geographical Information Systems) tabulations.
  • Needs further investigation.
Things we didn’t learn but still might

• What are tradeoffs between different disclosure limitation methods?
  • 1940 Census data application allows us to compare DP with other methods, e.g., swapping, more completely synthesized data, other approaches.
  • Providing more of these would be extremely useful!

• What about comparative assessment of disclosure avoidance methods with other sources of data distortion?
  • Would be useful to assess how DP noise-infusion compares with ways of dealing with missing data & other sources of data inaccuracies.
  • Here we won’t know “truth,” i.e., data without any distortion, but assessment would enrich our understanding of how much outcomes (resource allocation, segregation) are distorted by top-coding, assignment of missing data (Little & Rubin, 1987, 2003), etc.

• How robust are existing measures (statistics) to Noise-Infusion & are their more robust ones?
  • Two studies assess degree of segregation using existing Dissimilarity Index (Duncan & Duncan, 1955). There are many other measures of segregation (Reardon, 2006).
  • Would be useful to better understand & assess sensitivity of this and other segregation measures to noise-infusion.
  • More generally, need to better understand sensitivity of existing measures & statistics to noise-infusion.
Census Bureau’s Twin Objectives

Our Mission
The Census Bureau’s mission is to serve as the nation’s leading provider of quality data about its people and economy.

Disclosure Avoidance

Title 13, U.S. Code
The Census Bureau is bound by Title 13 of the United States Code. These laws not only provide authority for the work we do, but also provide strong protection for the information we collect from individuals and businesses.

Title 13 provides the following protections to individuals and businesses:

- Private information is never published. It is against the law to disclose or publish any private information that identifies an individual or business such, including names, addresses (including GPS coordinates), Social Security Numbers, and telephone numbers.
- The Census Bureau collects information to produce statistics that describe the population and improve the efficiency of the Federal government.
- Census Bureau employees are sworn to protect confidentiality; they are sworn to uphold Title 13 and are subject to severe penalties if they violate the law.
- Violating the law who violates this law is subject to a fine of up to $50,000.
The Public Policy Challenge

• We know where data users & data custodians are on this tradeoff.

• **Key question**: Where is “the public” on this tradeoff?

• **Another Question**: Can we improve this tradeoff by working on the “privacy front”?
On the “Privacy Front” 1

• Need to know where “the public” is on privacy vs data utility tradeoff.
  • “Privacy Paradox”: Surveys show privacy is primary concern, yet people readily reveal information on-line (Kokolakis, 2017).

• Important to know willingness of public to tradeoff disclosure risks for data utility.
  • Challenges to eliciting information about this willingness to accept greater disclosure risks in provision of data to government (Census Bureau).

• In addition, can we change public’s willingness to accept greater disclosure risk in return for forms of protection?
On the “Privacy Front” 2

• Problem similar to study subjects/respondents williness to consent to provide personally identifiable information, e.g., their income or their health records.

• Hotz & Slanchev (2017) assessed such willingness to consent (WTC) to participation in health care study (MURDOCK).
  • Subjects randomly assigned differing vignettes about data security assurances &/or research and personal benefits (data utility) of study.
  • Collected demographics, knowledge & use of privacy protection schemes & “trust” & “trust of institutions” (govt., health care)

Findings:

a) WTC did vary, although not much across demographic groups.
b) WTC not increased by providing additional data security assurances.
c) WTC higher among subjects with greater “trust,” in general and “trust in institutions” in particular.
d) WTC higher among subjects with greater knowledge of &/or utilization of privacy protection schemes.
The “Privacy Front”: Some Questions

- Can we do better job of educating the public about disclosure risks to providing data to Census and utility of these data?

- Would this change public’s willingness to provide more accurate (and “disclosive”) data to Census?

- Can we do things to build trust in institutions like the Census Bureau?
  - Won’t be easy! (see graphs)

- Can we provide the public with “insurance” or “protection” against consequences of disclosure rather than “guarantees” of privacy?

- While I’m not sanguine on any of above, seems worth additional research & exploration, given perceived disparity with what public wants & what data users want.
References


Differential Privacy for 2020 US Census (II)

Salil Vadhan
Harvard University
REMARKS ON DIFFERENTIAL PRIVACY FOR 2020 CENSUS

Salil Vadhan
Privacy Tools Project
Harvard University

HDSR Inaugural Symposium
October 25, 2019

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our funders.
1) These kinds of studies are important

- Learn how to make best use of DP releases.
- Inform the (unavoidable) decisions about which queries should be prioritized for accuracy.
- We need more of them!
2) The vulnerabilities are real

Releasing “too many” statistics that are “too accurate” necessarily makes one vulnerable to:

- **Database Reconstruction**: reconstructing almost the entire underlying dataset [Dinur-Nissim ’03,…]
  - Applied in practice to Census releases [Garfinkel-Abowd-Martindale `18] and Diffix [Cohen-Nissim `19].

- **Membership Inference**: determining whether a target individual is in the dataset [Dwork-Smith-Steinke-Ullman-V. `15]
  - Applied in practice to genomic data [Homer et al. `08,…] and ML as a service [Shokri et al. `17,…]
An example from industry: ML models memorize training data

[Reddit users `18], see also [Zhang et al. `17, Song et al. `17, Carlini et al. `19]
[slide based on one from Ilya Mironov]
An example from industry: ML models memorize training data

[Reddit users `18], see also [Zhang et al. `17, Song et al. `17, Carlini et al. `19]
[slide based on one from Ilya Mironov]
2) The vulnerabilities are real

Releasing “too many” statistics that are “too accurate” necessarily makes one vulnerable to:

- **Database Reconstruction**: reconstructing almost the entire underlying dataset [Dinur-Nissim `03,…]

- **Membership Inference**: determining whether a target individual is in the dataset [Dwork-Smith-Steinke-Ullman-V. `15]

Differential Privacy is only known framework for controlling accumulated risk.

Difficult policy choices are unavoidable.
3) Noise in Census data is not new

Past Census releases had noise from:
- Respondent mistakes
- Non-response
- Editing
- Imputation
- Traditional Disclosure Avoidance (e.g. swapping)
Log Difference Between IPUMS and NHGIS

Population

White

Blacks

[Asquith et al. 2019]
3) Noise in Census data is not new

Past Census releases had noise from:
- Respondent mistakes
- Non-response
- Editing
- Imputation
- Traditional Disclosure Avoidance (e.g. swapping, suppression)
  - Methods had to be kept secret, in contrast to DP.
  - Q: what would be effect on utility if it were done more aggressively to protect against attacks?
  - Chetty et al. (2018) find “count-based suppression can create bias in ways that cannot be easily identified or corrected ex-post.”
Slope = 0.136
(0.015)

Association between Teenage Birth and Two-Parent Share: Noise-Infused Data

Source: Chetty, Friedman, Hendren, Jones, Porter (2018)
Association between Teenage Birth and Two-Parent Share: Count-Suppressed Data

Slope = 0.028

Source: Chetty, Friedman, Hendren, Jones, Porter (2018)
4) Need to develop new statistical methods

- Standard “off-the-shelf” methods and metrics (e.g. dissimilarity index) may not be the best choices on DP data.

- Need to use/develop new methods that are robust to and account for noise.
  [Dwork-Lei `09, Vu-Slavkovic `09, …, Couch-Kazan-Shi-Bray-Groce `19, Alabi-McMillan-Sarathy-Smith-V. `19]
Differential Privacy for 2020 US Census (II)

Ruobin Gong
Rutgers University
Drawing the Right Inference:
Differential Privacy Enables Proper Uncertainty Quantification

Ruobin Gong

Department of Statistics
Rutgers University

Differential Privacy for 2020 US Census (II)
HDSR Inaugural Symposium · Oct 25, 2019
The studies by Brummet et al., Riper et al., and Asquith et al. supply valuable and timely insight on the potential impact of differential privacy (DP) as the disclosure avoidance system of the 2020 U.S. Census.

Inference based on differentially private Census data should come with proper uncertainty quantification:

1. beyond point estimates: intervals, distributions, and probabilistic statements;
2. incorporating DP mechanism as an additional layer of uncertainty in models.

Transparency enables accurate modeling of the DP mechanism. This is the most reliable way to ensure correctness in the resulting inference, when a (calculated) loss of statistical efficiency is present in the data.
Suppose a simple linear model between vector counts $x$ and $y$:

$$y = \beta_0 + \beta_1 x + e.$$ 

Ordinary least squares produce consistent estimators

$$\hat{\beta}_0 \to \beta_0, \quad \hat{\beta}_1 \to \beta_1.$$ 

Treating $(x, y)$ with $\epsilon$-DP mechanism

$$y_{dp} = y + w, \quad x_{dp} = x + z, \quad w, z \sim Lap\left(\epsilon^{-1}\right)$$

standard Normal and Laplace densities
Naïvely fitting the original model with differentially private data

\[ y_{dp} = \beta_0 + \beta_1 x_{dp} + e, \]

the resulting least squares estimates will miss the mark:

\[ \hat{\beta}_0^{dp} \approx \beta_0 + \alpha_{x,z} \beta_1, \quad \hat{\beta}_1^{dp} \approx \beta_1 - \gamma_{x,z} / \beta_1, \]

where

- \( \alpha_{x,z} \) has sign that depends on \( x \) and its DP mechanism;
- \( \gamma_{x,z} \in (0, 1) \): strength of association between \((x, y)\) is underestimated;
- both estimates suffer from inflated variance.

Therefore, uncertainty quantification according to the naïve fitting of the original model on DP data is not trustworthy.
Figure: Illustrated with $x_i \sim \text{Pois}(10)$, $y_i = -5 + 4x_i + e_i$, $e_i \sim N(0, 5^2)$, $n = 10$, at various levels of privacy budget $\epsilon$. Smaller $\epsilon$ induces more misguided confidence regions.
Quantifying uncertainty: the correct approach

If the original model is adequate for the ideal data \( \mathcal{D} = (x, y) \), naively adapting it to the privatized data \( \mathcal{D}_{dp} = (x_{dp}, y_{dp}) \) is very likely inadequate:

\[
\begin{align*}
  y &= \beta_0 + \beta_1 x + e \\
  \implies y_{dp} &= \beta_0 + \beta_1 x_{dp} + e
\end{align*}
\]

The correct approach **augments** the original model with the DP mechanism:

\[
\begin{align*}
  (y_{dp} - w) &= \beta_0 + \beta_1 (x_{dp} - z) + e, \quad w, z \sim \text{Lap}(\epsilon^{-1})
\end{align*}
\]

**A General Construction**

Likelihood for \( \beta \) based on privatized data \( \mathcal{D}_{dp} \) (observed) is integrated over the ideal data \( \mathcal{D} \) (missing), with respect to the DP mechanism:

\[
L(\beta; \mathcal{D}_{dp}) = \int \eta(\mathcal{D}_{dp} | \mathcal{D}) \ f(\mathcal{D} | \beta) \, \partial\mathcal{D}
\]

**Transparency of the DP mechanism enables accurate modeling.**
Figure: Correct model (green) fitted via Monte Carlo EM (G. 2019) vs. naïve model (gray) on six instances of DP protected datasets ($\varepsilon = 0.2$). Displayed 95% confidence ellipses are based on normal approximations at the MLE.
Select Literature

- **Missing data: theory, practice, and computation**
  - DP as an MCAR or MAR mechanism \((\text{Little & Rubin, 2002})\)
  - Inference under input uncongeniality \((\text{Meng, 1994; Xie & Meng, 2017})\)
  - Quantification of fractional missing information \((\text{Louis, 1982; Meilijson, 1989})\)
  - Expectation-maximization, data augmentation, multiple imputation \((\text{Dempster et al., 1977; Tanner & Wong, 1987; Wei & Tanner, 1990; Park et al., 2017})\)
  - Approximate Bayesian computation and exact inference for DP data \((\text{Fearnhead & Prangle, 2012; Gong, 2019})\)

- **Hierarchical modeling**
  - Random effects, generalized linear mixed models, latent variable and hidden Markov models \((\text{McCulloch, 1997; Booth & Hobert, 1999; Sinha, 2004})\)
  - Model misspecification \((\text{Miller & Dunson, 2018})\)
  - Bayesian model averaging \((\text{Hoeting et al., 1999})\)

- **Robust statistical estimation**
  - Huber & Ronchetti (2009); Dwork & Lei (2009); Lei (2011); Avella-Medina (2018)


Differential Privacy for 2020 US Census (II)

Mark Hansen
Columbia University
Discussion for the Harvard Data Science Review
Inaugural Symposium, October 25, 2019

Mark Hansen
Columbia Journalism School
We set out to make the operations of the Bureau’s Differential Privacy scheme understandable by various stakeholders — from civil society groups to the general public.

Our approach was to work with the code provided by the Bureau for 1940, creating many additional simulations for each value of the Privacy Loss Budget.

The result lets users assess the variation across various geographies, deciding if the resulting tabulations are too noisy for their needs. Our hope is that this approach might help inform how to involve more people in the process of setting the Privacy Loss Budget.

The entire United States in 1940 seemed overwhelming to us and so Timothy Jones, Alex Calderwood and I sought to skinny it down a little. We were fortunate enough to be able to get some assistance on at least a corner of the US for that year — John Logan of the Urban Transition HGIS project gave us shape files for LA’s enumeration districts.

Mind you, the map is still noisy and we have work to do.
Uncertainty across replicates for two values of %white, plots describing different values for the loss budget.
This is the city of Los Angeles — some of the grey areas are cities that are part of the county, some are orphaned by the complex process to match up the hand-written maps. Given small-scale geography, we can start to ask questions like whether the “polish” added to the top-down approach that assures tables add up across various levels of geography has any spatial impacts. At very least we can see whether maps drawn at small levels of geography tell stories that we know to be true from other sources — It’s 1940 LA and we have redlining, etc.
Simulations for tracts associated with Venice, CA
Thank you

Mark Hansen
Columbia Journalism School
Inaugural Symposium

hdsr.mitpress.mit.edu

Differential Privacy for 2020 US Census (II)

Paul Ohm
Georgetown University