The Challenge of Scientific Reproducibility and Privacy Protection for Statistical Agencies

John M. Abowd
Associate Director for Research and Methodology and Chief Scientist,
U.S. Census Bureau

Census Scientific Advisory Committee
September 15, 2016

Acknowledgments and Disclaimer

- Parts of this talk were supported by the National Science Foundation, the Sloan Foundation, and the Census Bureau (before and after my appointment started)
- The opinions expressed in this talk are my own
- All tables and graphics based on internal Census Bureau data were reviewed to insure that no confidential data were disclosed

Outline

- Use the confidential data
- Respect a privacy-loss budget
- Prove that the rate of privacy loss in published data is consistent with the budget
- Prove that the released data are robust to all future attacks
- The future is managed access to the confidential data with privacy-preserving statistical analyses, with examples

1. Use the Confidential Data

- To the maximum extent possible, estimation and inference should be based on the original confidential data
- This is where current SDL methods fail the scientific community most egregiously
  - Suppression creates massive nonignorable missing data biases when used in detail tables
  - Almost all other methods disclose insufficient information to correct the inferences

See: Abowd and Schmutte (2015) "Economic Analysis and Statistical Disclosure Limitation" Brookings Papers on Economic Activity (Spring): 221-293, media/Projects/BPEA/Spring-2015-Revised/AbowdText.pdf?la=en
http://www.brookings.edu/~/media/Projects/BPEA/Spring-2015-Revised/AbowdText.pdf?la=en

Inference-valid SDL

- Two SDL methods that do perform all of the scientific calculations on the original confidential data and allow provably valid inference:
  - Synthetic data with validation
  - Privacy-preserving data analysis
- These are the scalable long-term components of modernized SDL
- Access via enclaves like the Federal Statistical RDCs and virtual enclaves is also part of the solution, but it doesn’t scale as well, and the SDL remains problematic

2. Respect a Privacy-loss Budget

- There are provable limits to the amount of information that can be published from any given confidential inputs
- These limits imply that there is a budget that relates published data quality to privacy loss
- Data publication always implies some privacy loss
- Respecting a privacy-loss budget means quantifying the privacy loss from each publication

Example: Randomized Response

- Randomized response (asking a survey respondent one of two questions at random: one sensitive (SQ), one innocuous; the innocuous question can also be randomized, if needed) is provably privacy-loss protective
- Privacy loss is bounded by the maximum Bayes factor designed into the randomized response protocol

$$\max BF = \frac{\frac{Pr[SQ = Yes|A = Yes]}{Pr[SQ = No|A = Yes]}}{\frac{Pr[SQ = Yes]}{Pr[SQ = No]}} = \frac{Pr[A = Yes|SQ = Yes]}{Pr[A = Yes|SQ = No]} = \frac{(1/2) + (1 - 1/2)\,1/2}{(1 - 1/2)\,1/2} = 3$$

- In privacy-preserving data analysis this bound is stated as the logarithm of the maximum Bayes factor
- If the sensitive question is asked with probability ½ and the innocuous question is “yes” with probability ½, then the maximum Bayes factor is 3, and ln 3 = 1.1
- The privacy-loss expenditure (ε-differential privacy) is 1.1 (for all possible input databases and all future attacks of any form)

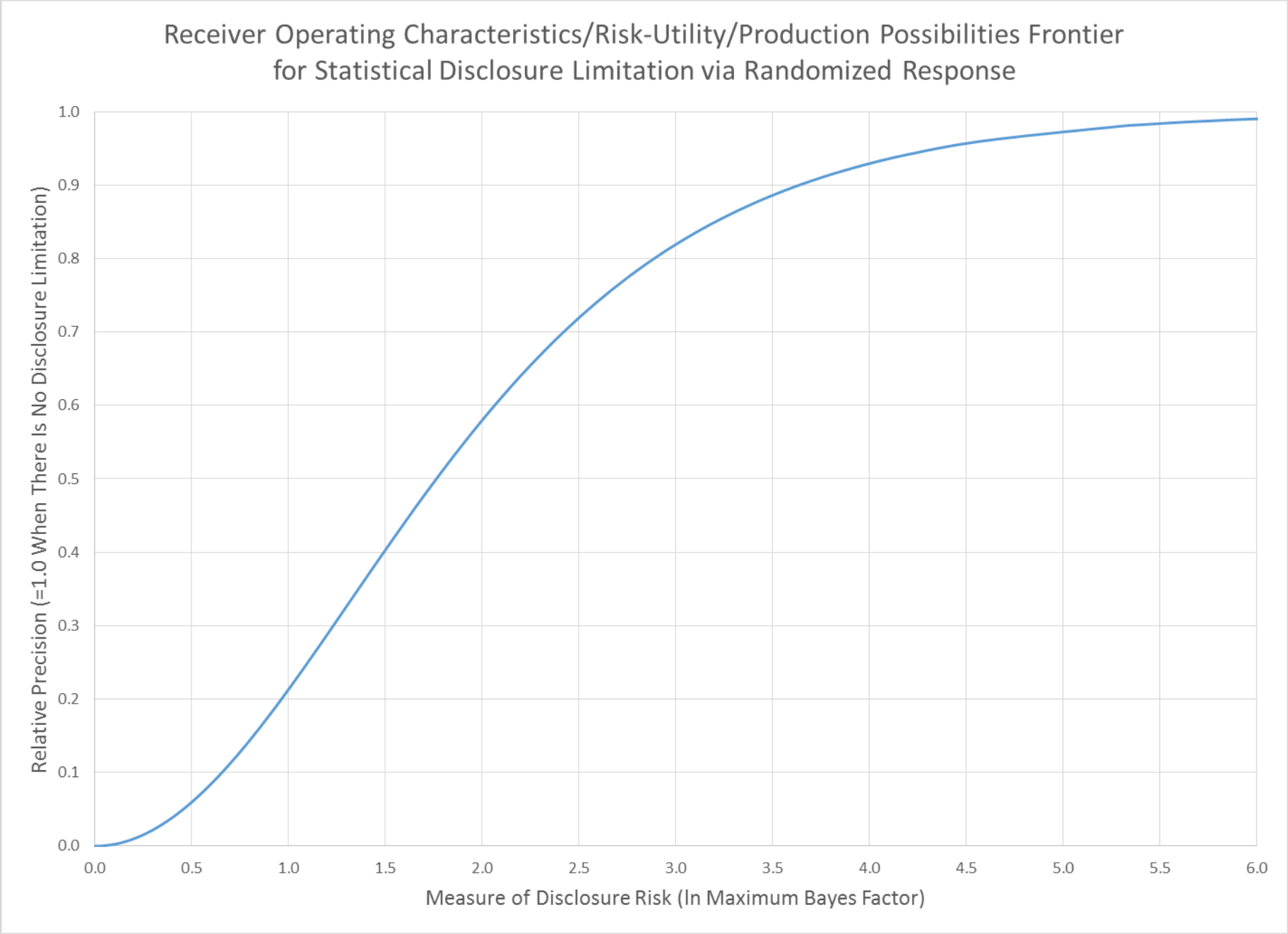
Randomized response: Warner, Stanley L. (1965) and Greenberg , Bernard G., Abdel-Latif A. Abul-Ela, Walt R. Simmons and Daniel G. Horvitz (1969). SDL uses: Fienberg and Steele (1998), Du and Zhan (2003) and Erlingsson , Úlfar, Vasyl Pihur,and Aleksandra Korolova (2014)

What Happens to Data Quality?

- The sampling precision (inverse of the sampling variance) associated with the randomized-response estimator is proportional to the sampling precision when all respondents are asked the sensitive question

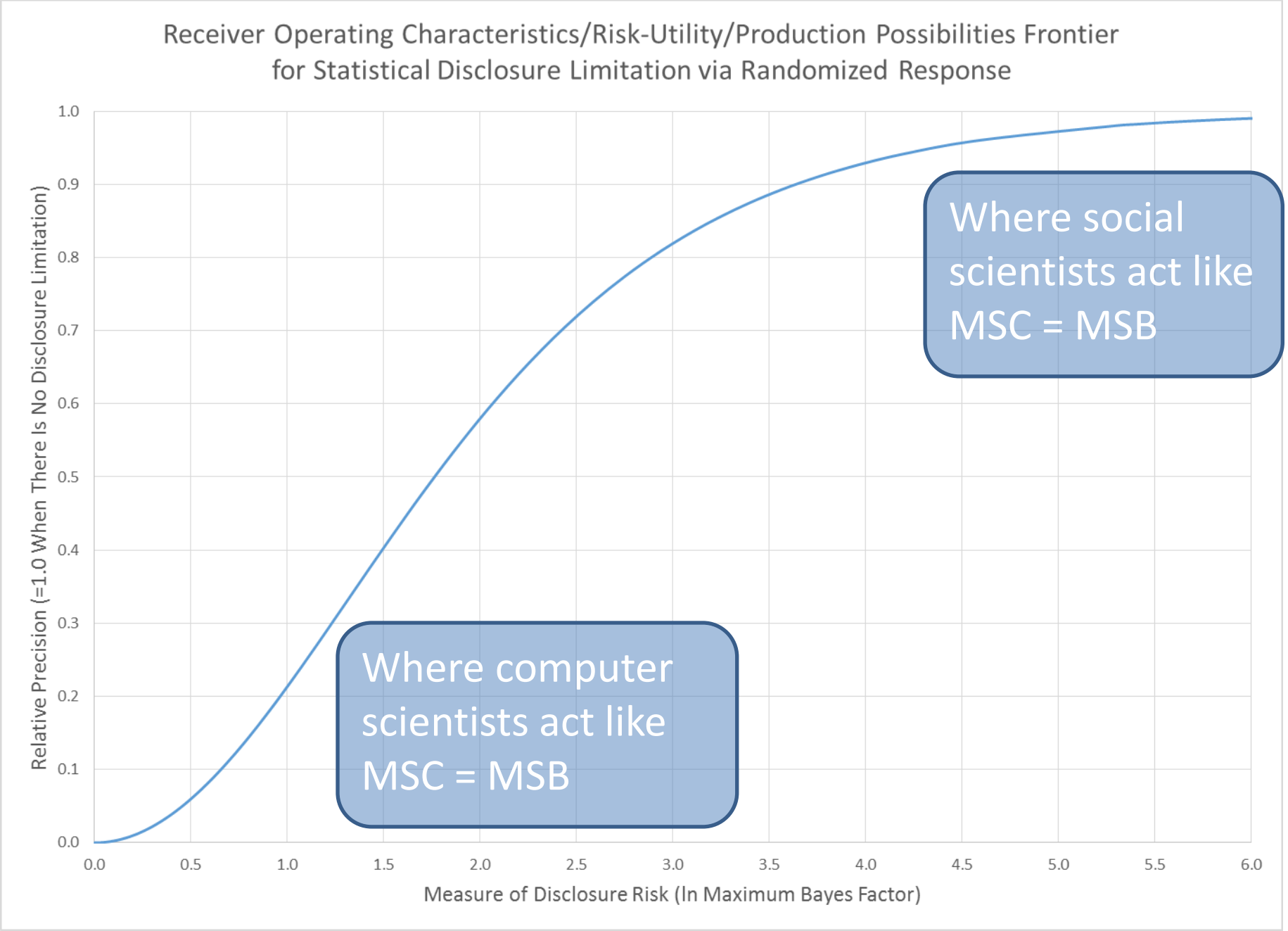
$$Rel.Precision = \frac{\{Pr[Ask\ Sensitive\ Q]\}^2 \frac{n}{\theta(1-\theta)}}{\frac{n}{\theta(1-\theta)}} = \left\{\frac{1}{2}\right\}^2 = 0.25$$

- Under the same conditions in which the privacy loss is ln 3, the relative sampling precision is 25% of the most accurate estimator
- The graph shows this trade-off from privacy-loss expenditures of zero to 6 (ε-differential privacy, x-axis)
- Data quality is measured by the relative precision (y-axis)


Receiver Operating Characteristics/Risk-Utility/Production Possibilities Frontier
for Statistical Disclosure Limitation via Randomized Response
Relative Precision (=1.0 When There Is No Disclosure Limitation)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Measure of Disclosure Risk (ln Maximum Bayes Factor)

SDL is Technology

- Properly designed methods operate with a privacy-loss budget constraint
- The price of increasing data quality (more accurate publications) in terms of increased privacy loss is the slope of the technology frontier:
  - Economics: Production Possibilities Frontier
  - Forecasting models: Receiver Operating Characteristics Curve
  - SDL: Risk-Utility Curve (with risk on the x-axis)
- All exactly the same thing
- None able to select an optimal point


Receiver Operating Characteristics/Risk-Utility/Production Possibilities Frontier
for Statistical Disclosure Limitation via Randomized Response
Relative Precision (=1.0 When There Is No Disclosure Limitation)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Measure of Disclosure Risk (ln Maximum Bayes Factor)
Where social scientists act like MSC = MSB
Where computer scientists act like MSC = MSB

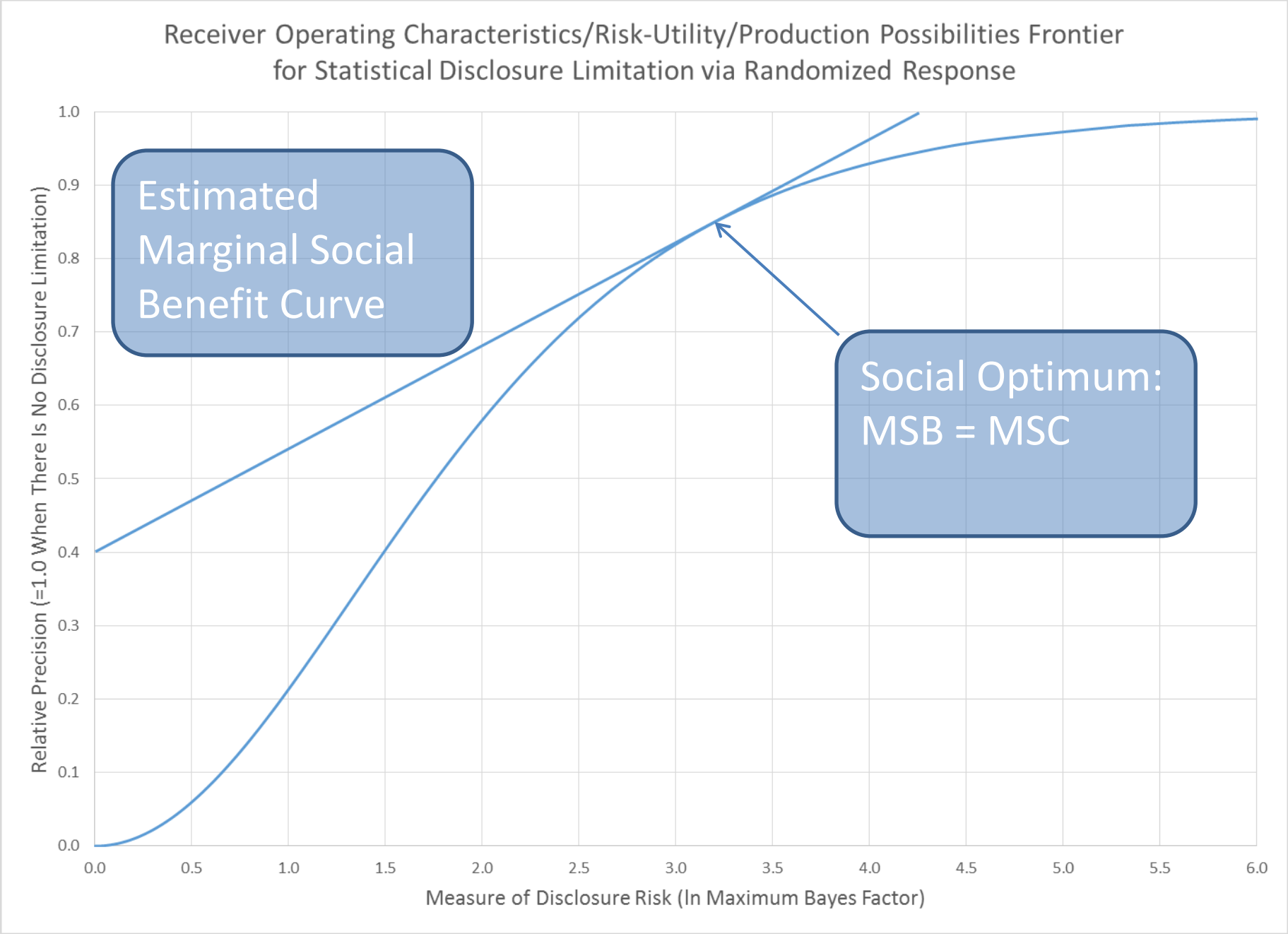
Some Examples

- Dwork (2008): "The parameter ε in Definition 1 is public. The choice of ε is essentially a social question and is beyond the scope of this paper. That said, we tend to think of ε as, say, 0.01, 0.1, or in some cases, ln 2 or ln 3. If the probability that some bad event will occur is very small, it might be tolerable to increase it by such factors as 2 or 3, while if the probability is already felt to be close to unacceptable, then an increase by a factor of $e^{0.01} \approx 1.01$ might be tolerable, while an increase of e, or even only $e^{0.1}$, would be intolerable."
- Dwork (2011): "The parameter ε is public, and its selection is a social question. We tend to think of ε as, say, 0.01, 0.1, or in some cases, ln 2 or ln 3."
- In OnTheMap, $\varepsilon = 8.9$, which was required to produce tract-level estimates with acceptable accuracy

How to Think about the Social Choice Problem

- The marginal social benefit from increased data quality is the sum of all citizens' willingness-to-pay for data quality with increased privacy loss
- Can be estimated from survey data
- The next slide shows how to use the estimate from survey data to select an optimal (data quality, privacy loss) pair

See Abowd and Schmutte (2015) "Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods" http://digitalcommons.ilr.cornell.edu/ldi/22/ (under review)


Receiver Operating Characteristics/Risk-Utility/Production Possibilities Frontier
for Statistical Disclosure Limitation via Randomized Response
Relative Precision (=1.0 When There Is No Disclosure Limitation)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Measure of Disclosure Risk (ln Maximum Bayes Factor)
Estimated Marginal Social Benefit Curve
Social Optimum: MSB = MSC

What's the Right Choice?

- Combining the survey data responses to estimate marginal social benefit with the technology associated with randomized response
- Ask the sensitive question 92% of the time
- Publish data with 85% of the sampling precision of the estimator that is not privacy-preserving
- Note that the "sensitive" question can be anything, and is usually already recorded in the confidential database
- In this case, randomized response is a form a post-randomization SDL where all parameters are public
- More examples in Abowd and Schmutte (2015)

3. Prove That the Privacy-loss Budget Was Respected

- You can’t respect a privacy-loss budget without quantifying the expenditure of each publication
- The collection of the algorithms taken altogether must satisfy the privacy loss budget

The Database Reconstruction Theorem

- Also called the Fundamental Law of Information Reconstruction
- Dinur-Nissim (2003) database reconstruction theorem: Any finite confidential database can be reconstructed to arbitrary accuracy from a finite sequence of queries
- Many agencies act as if this result does not apply to their SDL
- But agencies should acknowledge that it does, especially if suppression is the primary SDL
- Holan et al. (2010) show how vulnerable some databases, which use only suppression and publish unaltered cells, can be to exact reconstruction

4. Prove That the Algorithms Are Resistant to All Future Attacks

- The information environment is changing much faster than it was when most current SDL methods were invented
- It may no longer be reasonable to assert that a product is empirically safe given best-practice SDL prior to its release
- In the randomized response example, the privacy loss of ln 3 never grows regardless of any future data publications from external sources
- This result holds for the confidentiality protections in any formally private system

A Long Row to Hoe

- It's going to take a concerted research and engineering effort to bring SDL into the 21st century
- But the scientific integrity of agency enterprises requires that we tackle this challenge
- The first step is experimentation with the technologies known to work:
  - Synthetic data with validation
  - Privacy-preserving data analysis via online query systems

Working Current Examples

- OnTheMap (residential side adding employer side)
- SIPP Synthetic Data server with validation
- LBD Synthetic Data server with validation
- RAPPOR

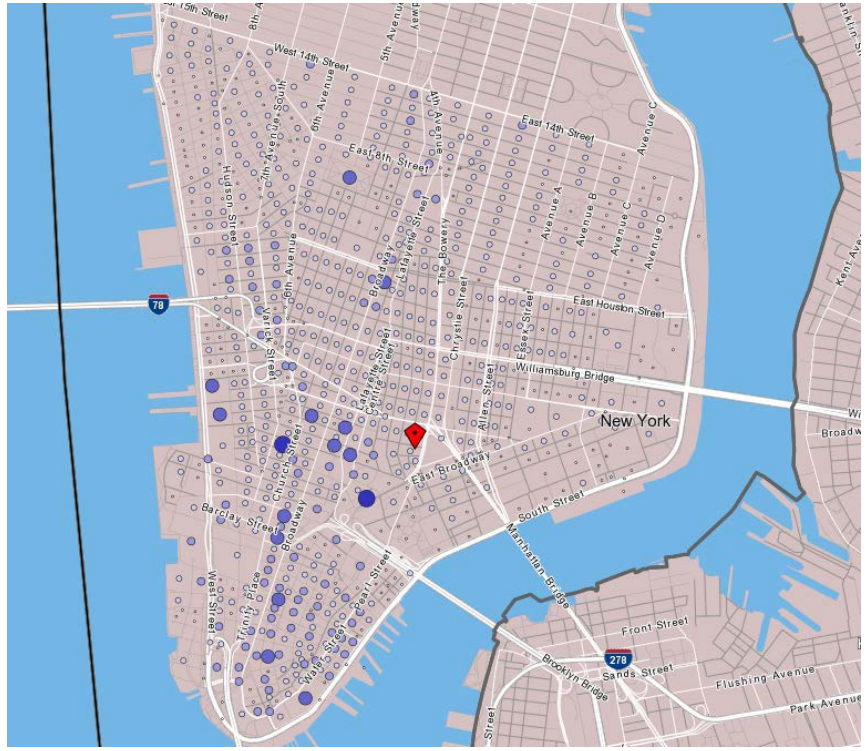
Example 1: OnTheMap Employer Data

Background/Motivation

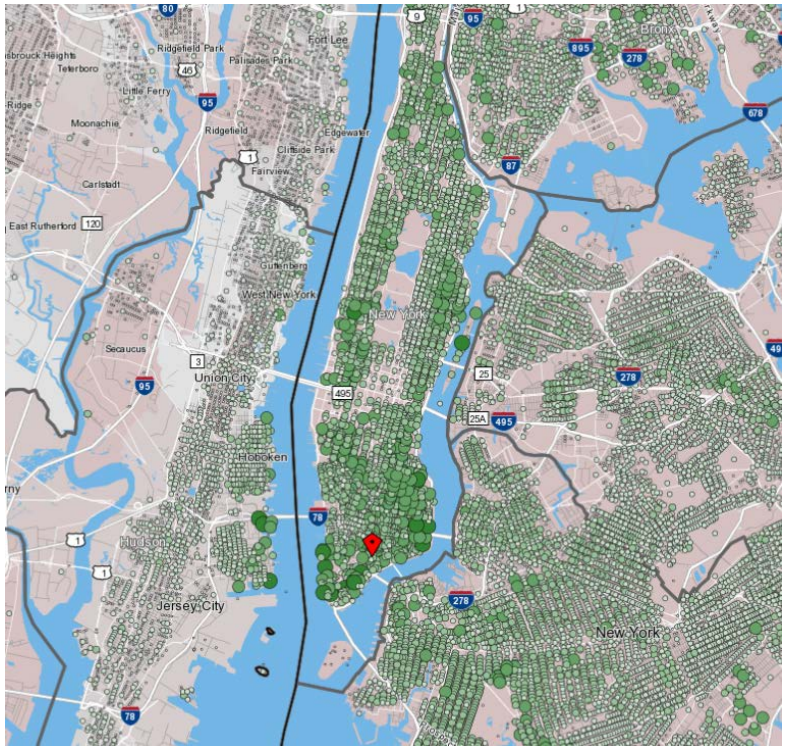
- Currently, the Census Bureau releases the LEHD Origin-Destination Employment Statistics (LODES).
- In protecting confidentiality of workers and firms, LODES utilizes:
  - Permanent multiplicative noise distortion factors at the employer and establishment level (plus synthetic data methods for small cells) for employment counts.
  - Synthetic data methods using probabilistic differential privacy for residential location.
- Research goal: Develop a formal provably-private protection method based on differential privacy to replace the noise distortion factors.

Existing LODES Data in OnTheMap Application

Employment in Lower Manhattan



Residences of Workers Employed in Lower Manhattan

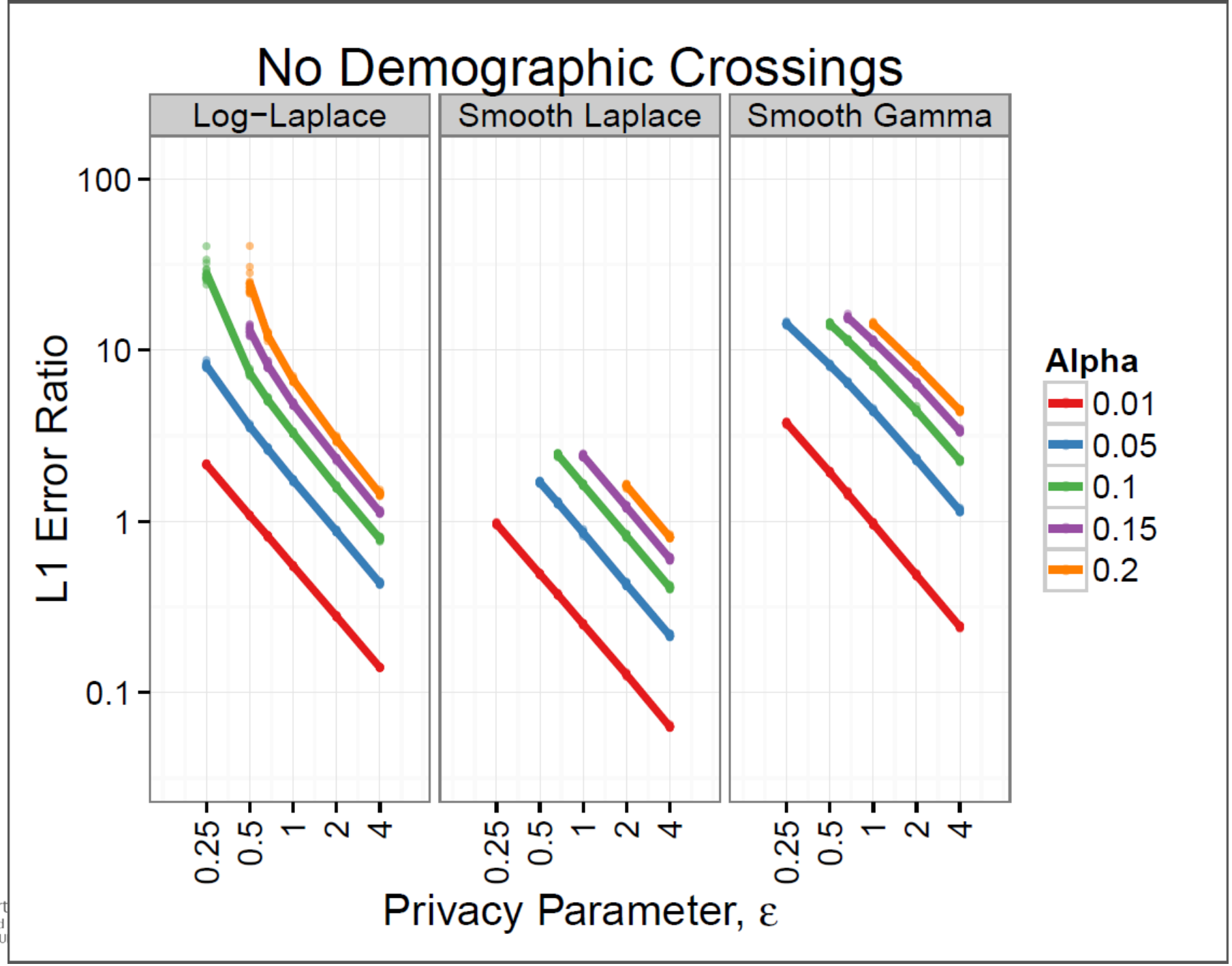


Available at http://onthemap.ces.census.gov/

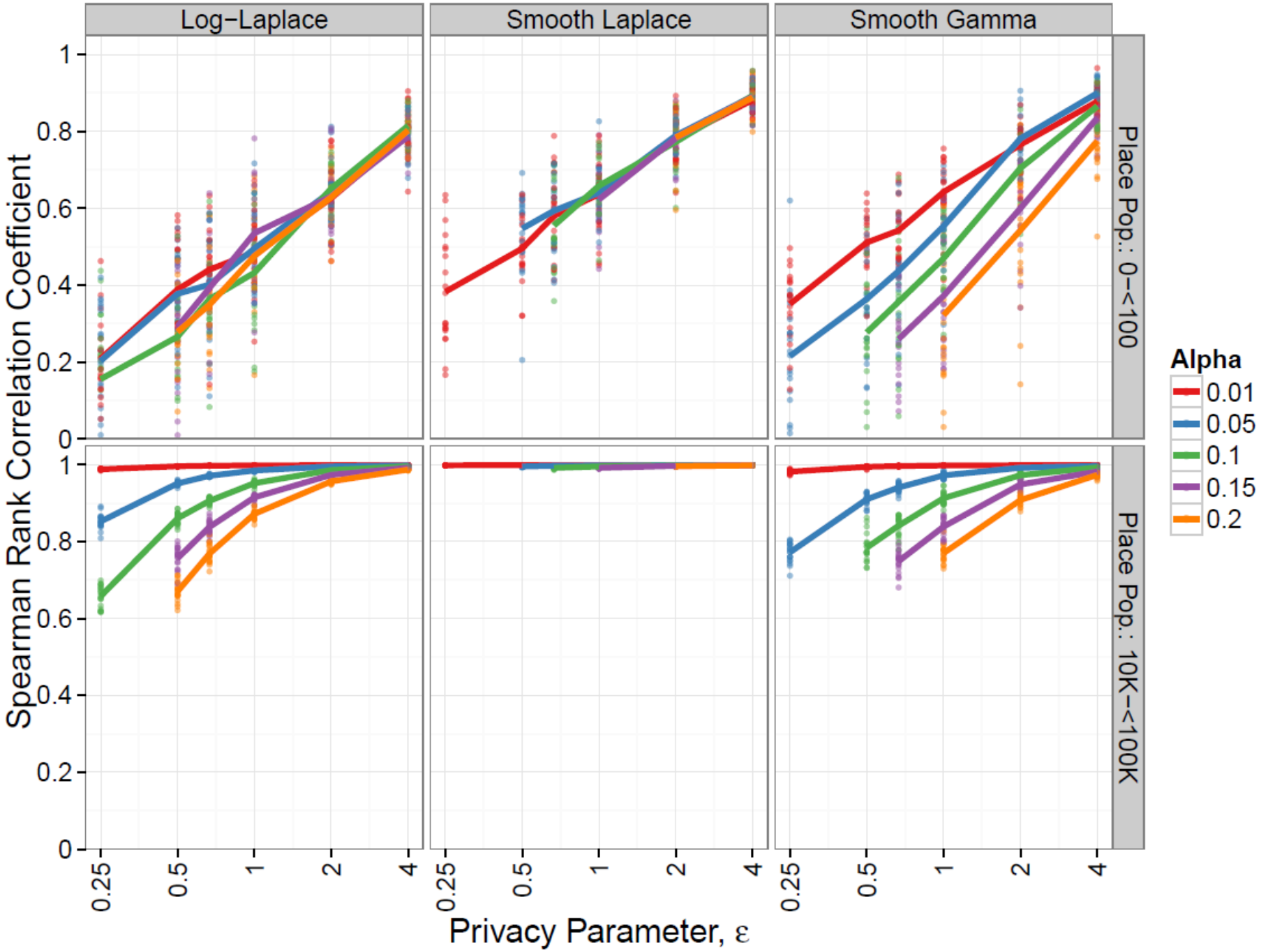
Goals of New Protection System

- It should **answer marginal queries** (employment counts) over protected characteristics (e.g., sex, age, race)
- It should use algorithms that have **provable privacy guarantees** for both **individuals and employer businesses**
- It should **perform comparably** to the existing system in terms of **data quality**

Joint work with Machanavajjhala, Haney, Graham, Kutzbach and Vilhuber (in progress)



United States Census Bureau | U.S. Depart
Economics and
U.S. CENSUS BU
census.gov



United States Census Bureau | U.S. D
Economi
U.S. CEN
census.g

Example 2: SIPP Synthetic Data

Adapted from Abowd and Vilhuber "Understanding Social and Economic Data" (Cornell INFO SCIENCE 7470 https://www.vrdc.cornell.edu/info747x/) ©2016, all rights reserved

Survey of Income and Program Participation

About this Survey
Information for Respondents
Data
Events
Guidance for Data Users
DataFerrett
SIPP FTP site
Synthetic SIPP Data
SIPP Users' Guide
Methodology
News
Publications
Technical Documentation
Working Papers

Synthetic SIPP Data

Tweet Share

Background on the SIPP Synthetic Beta

The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receip and were produced by Census Bureau staff economists and statisticians in collaboration with researchers at Cornell University, the SSA anc the IRS. The purpose of the SSB is to provide access to linked data that are usually not publically available due to confidentiality concerns. To overcome these concerns, Census synthesizes, or models, all the variables in a way that changes the record of each individual so as to preserve the underlying covariate relationships between the variables. Only gender and a link to the first reported marital partner are not altered by the synthesis process and still contain their original values.

Nine SIPP panels (1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, and 2008) form the basis for the SSB, with a subset of variables available across all the panels selected for inclusion and harmonization of variable definitions across the years covered by the panels. Administrative data are added and some editing is done to correct for logical inconsistencies in the IRS and Social Security earnings and benefits data. Thus, the SSB is a particularly appealing data set for new SIPP users because little data preparation is needed. A complete list of variables included in SSB version 6.0, along with details about the harmonization and editing, is available in our Codebook.

As part of the synthesis process, missing survey data and missing administrative data were multiply imputed. The resulting data sets are called the Completed Gold Standard Files and contain all original, non-missing, confidential values and imputed values in place of originally missing data. These files form the basis for evaluating results from the synthetic data. The goal of the SSB is to produce results that are qualitatively the same as results from the Completed Gold Standard Files.

Basic Structure of the SSB

- Census Bureau: Survey of Income and Program Participation
  - Core set of SIPP drawn from eight SIPP panels (1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, 2008)
  - All missing data items (except for structurally missing) are marked for imputation
- Internal Revenue Service: Tax Information Filings
  - Maintained at SSA, but derived from IRS records
  - Master summary earnings records (Summary Earnings Record, SER, topcoded at tax maximum income)
  - Master detailed earnings records (Detailed Earnings Record, DER, every statutory employer, no topcoding)

Basic Structure of the SSB

- Social Security Administration (SSA)
  - Master Beneficiary Record (MBR): applications and benefit determinations
  - Supplemental Security Record (SSR): supplemental insurance
  - 831 Disability File (F831): disability insurance
  - Payment History Update System (PHUS): actual payments
- Census-improve Administrative Data (source: SSA)
  - Numident: administrative birth and death dates
- All files combined using verified SSNs

⇒"Gold Standard" completed using multiple imputation for missing data

⇒(better name: Consolidated Harmonized SIPP Panels 1984-2008 with linked IRS/SSA data)

Basic Structure of SSB V5.1

- Couple-level linkage: the first person to whom the SIPP respondent was married during the time period covered by the SIPP panel
- SIPP variables only appear in years appropriate for the panel indicated by the PANEL variable
- Additional administrative data were added
  - some editing for logical inconsistencies in the IRS/SSA earnings and benefits data.

Basic Structure of SSB V6.0.2

- Same basic structure as V5.1
- Updated administrative data (through 2012)
- Panels for 1984 and 2008 added
- Many monthly benefit indicators from SIPP
- Extensive additional date variables (application and benefit dates)
- Base weight from SIPP
- Improved fertility history

SSB Documentation

- Overview: http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/DRBMemoTablesVersion2SSBv6_0.pdf
- Codebook: https://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssb/

Missing Values in the Gold Standard

- Values may be missing due to
  - [survey] Non-response
  - [survey] Question not being asked in a particular panel
  - [admin] Failure to link to administrative record (non-validated SSN)
  - [both] Structural missing (*e.g.*, income of spouse if not married)
- All missing values except structural are part of the missing data imputation phase of SSB

Scope of the Synthesis

- Never missing and not synthesized
  - gender
  - Link to spouse
- All other variables in the public use file were synthesized

Public Use of the SIPP Synthetic Beta

- Full version (16 implicates) released to the Cornell VirtualRDC Synthetic Data Server (SDS)
- Any researcher may use these data
- During the testing phase, all analyses must be performed on the Virtual RDC
- Census Bureau research team will run the same analysis on the completed confidential data
- Results of the comparison will be released to the researcher, Census Bureau, SSA, and IRS (after traditional disclosure avoidance analysis of the runs on the confidential data)

ASA News

AMERICAN STATISTICAL ASSOCIATION

Promoting the Practice and Profession of Statistics®

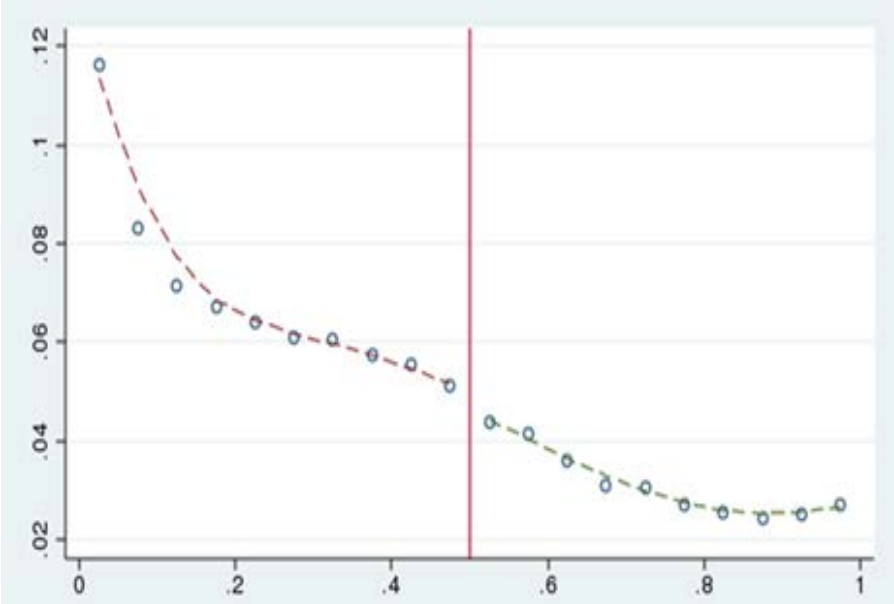
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

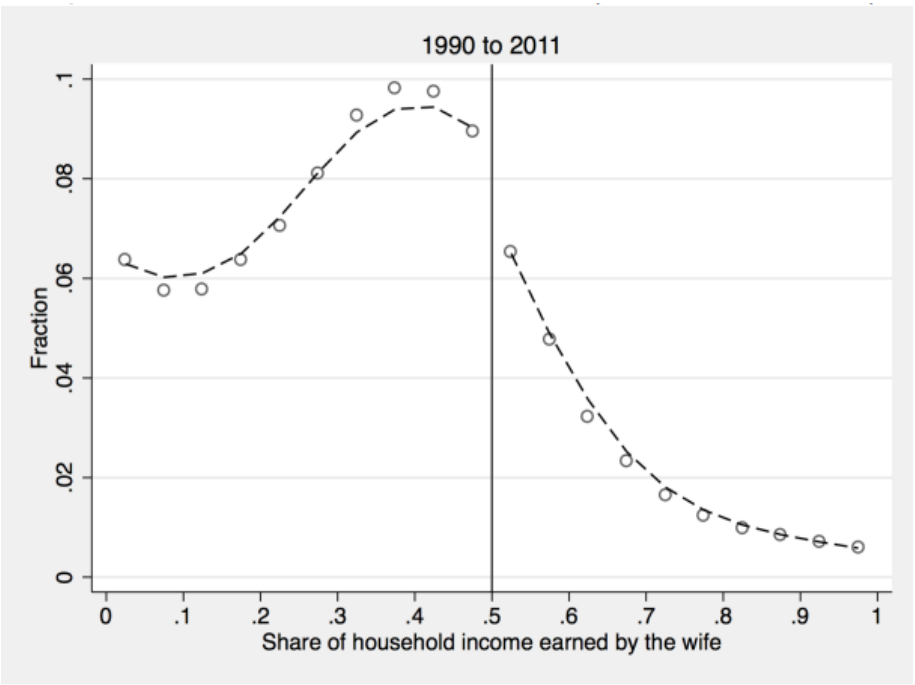
AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*





Bertrand, Kamenica and Pan (QJE 2015), doi: 10.1093/qje/qjv001

Example 3: Synthetic Longitudinal Business Database

Adapted from Abowd and Vilhuber "Understanding Social and Economic Data" (Cornell INFO SCIENCE 7470 https://www.vrdc.cornell.edu/info747x/) ©2016, all rights reserved

The Synthetic Longitudinal Business Database

Based on presentations by Kinney/Reiter/Jarmin/Miranda/Reznek[2]/Abowd

on July 31, 2009 at the

Census-NSF-IRS Synthetic Data Workshop

[link] [link]

Kinney/Reiter/Jarmin/Miranda/Reznek/Abowd (2011) "Towards Unrestricted Public Use Microdata: The Synthetic Longitudinal Business Database.", CES-WP-11-04

Work on the Synthetic LBD was supported by NSF Grant ITR-0427889, and ongoing work is supported by the Census Bureau. A portion of this work was conducted by Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Census Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau. Results have been screened to ensure that no confidential data are revealed.

Elements

(Economic Surveys and Censuses)
Issue: (item) non-response
Solution: LBD

(Business Register)
Issue: inexact link records
Solution: LBD

Match-merged and completed *complex integrated data*
Issue: too much detail leads to disclosure issue
Solution: Synthetic LBD

Public-use data
With novel detail

Novel analysis using Public-use data with novel detail
Issue: are the results right
Solution: Early release/SDS

Version of the LBD Used for Synthesis

- Economic census covering nearly all private non-farm business establishments with paid employees
  - Contains: Annual payroll and Mar 12 employment (1976-2005), SIC/NAICS, Geography (down to county), Entry year, Exit year, Firm structure
- Used for looking at business dynamics, job flows, market volatility, international comparisons

Longitudinal Business Database (LBD)

- Detailed description in Jarmin and Miranda
- Developed as a research dataset by the U.S. Census Bureau Center for Economic Studies
- Constructed by linking annual snapshot of the Census Bureau's Business Register

Longitudinal Business Database II

- Center for Economic Studies constructed
- Longitudinal linkages (using probabilistic record linking)
- Re-timed multi-unit births and
- Edits and imputations for missing data

Access to the LBD

- Different levels of access
- Public use tabulations – Business Dynamics Statistics http://www.ces.census.gov/index.php/bds
- "Gold Standard" confidential micro-data available through the Federal Statistical Research Data Center (FSRDC) Network
  - Most used dataset in the FSRDCs

Bridge between the Two

- Synthetic data set
  - Available outside the FSRDC
  - Providing as much analytical validity as consistent with confidentiality protections
  - Reduce the number of requests for special tabulations
  - Aid users requiring FSRDC access
- Experiment in public-use business micro-data

Why Synthetic Data with Validation

- Concerns about confidentiality protection for census of establishments
  - LBD is a test case for business data
- Criteria given for public release
  - No actual values of confidential values could be released
  - Should provide valid inferences while protecting confidentiality
  - All analyses are eligible for validation, user may publish output from the confidential data subjected to conventional SDL

Example 4: RAPPOR

RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response

Úlfar Erlingsson
Google, Inc.
ulfar@google.com

Vasyl Pihur
Google, Inc.
vpihur@google.com

Aleksandra Korolova
University of Southern California
korolova@usc.edu

ABSTRACT

Randomized Aggregatable Privacy-Preserving Ordinal Response, or RAPPOR, is a technology for crowdsourcing statistics from end-user client software, anonymously, with strong privacy guarantees. In short, RAPPORs allow the forest of client data to be studied, without permitting the possibility of looking at individual trees. By applying randomized response in a novel manner, RAPPOR provides the mechanisms for such collection as well as for efficient, high-utility analysis of the collected data. In particular, RAPPOR permits statistics to be collected on the population of client-side strings with strong privacy guarantees for each client, and without linkability of their reports.

This paper describes and motivates RAPPOR, details its differential-privacy and utility guarantees, discusses its practical deployment and properties in the face of different attack models, and, finally, gives results of its application to both

asked to flip a fair coin, in secret, and answer "Yes" if it comes up heads, but tell the truth otherwise (if the coin comes up tails). Using this procedure, each respondent retains very strong deniability for any "Yes" answers, since such answers are most likely attributable to the coin coming up heads; as a refinement, respondents can also choose the untruthful answer by flipping another coin in secret, and get strong deniability for both "Yes" and "No" answers.

Surveys relying on randomized response enable easy computations of accurate population statistics while preserving the privacy of the individuals. Assuming absolute compliance with the randomization protocol (an assumption that may not hold for human subjects, and can even be non-trivial for algorithmic implementations [23]), it is easy to see that in a case where both "Yes" and "No" answers can be denied (flipping two fair coins), the true number of "Yes" answers can be accurately estimated by $2(Y - 0.25)$, where

Thank you.

john.maron.abowd@census.gov