Research Data Centers, Reproducible Science, and Confidentiality Protection: The Role of the 21st Century Statistical Agency

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

Presentation to Summer DemSem
Sponsored by the Wisconsin Federal Statistical RDC
June 5, 2017

Acknowledgments and Disclaimer

- This talk is partially based on my 2016 Shiskin Lecture "How Will Statistical Agencies Operate When All Data Are Private?" and my 2016 FCSM Policy Conference talk "Why Statistical Agencies Need to Take Privacy-loss Budgets Seriously, and What It Means When They Do"
- I am particularly indebted to Lars Vilhuber, who taught me much of what I know about reproducible science
- 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 and not necessarily those of the Census Bureau or other research sponsors

Outline

- Reproducible science and formal privacy protection are joined at the hip
- Can we solve the p-hacking problem?
- Why does the database reconstruction theorem matter?
- What is a privacy-loss budget?
- How do you respect a privacy-loss budget?
- How do you prove that the rate of privacy loss in published data is consistent with the budget?
- What does it mean to prove that the released data are robust to all future attacks?
- Now, back to reproducible science, p-hacking and best practices

Reproducible Science and Formal Privacy Protection Are Joined at the Hip

- Reproducible science:
  - Provenance control and certification
  - Output verification from certified inputs
  - Controlling the generalizability/inference validity of conclusions
  - Archiving
  - Curation of data and metadata
- Formal privacy protection:
  - A confidential database contains a finite amount of information
  - Every published use exposes some of this information
  - Global privacy loss must be quantified
  - Once quantified, it is a public-policy decision how to manage it

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

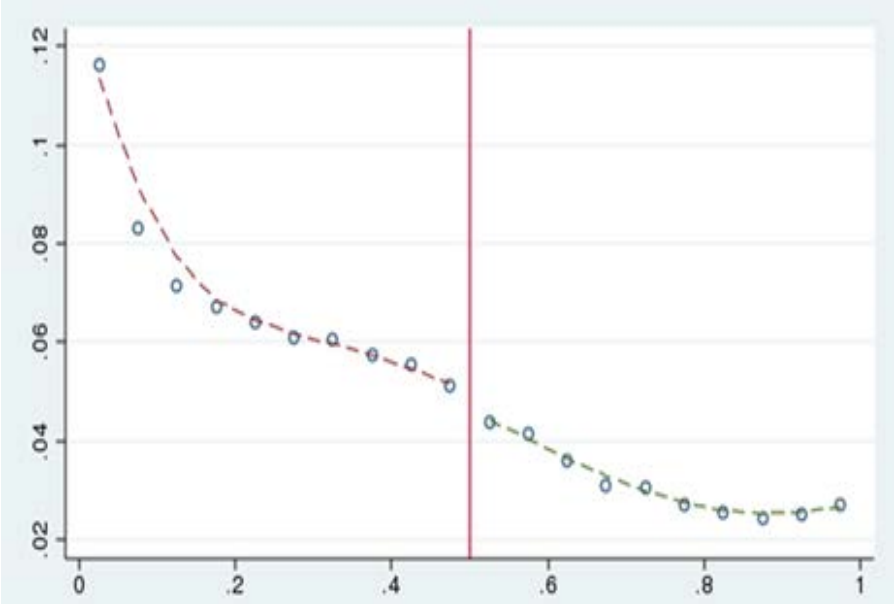
AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

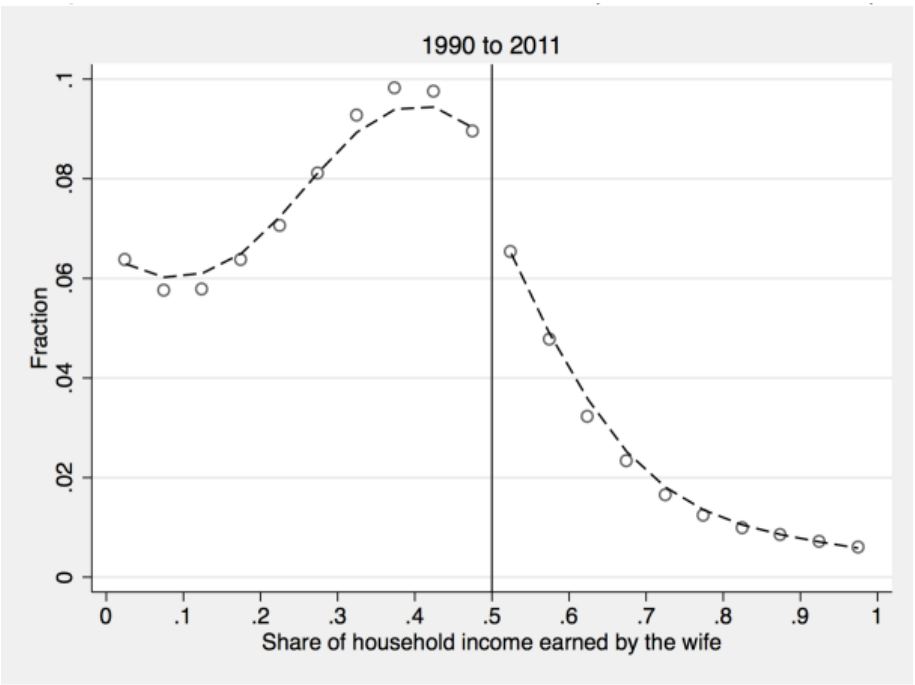
March 7, 2016

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.*

Synthetic SIPP



Confidential SIPP



Source: Bertrand, Marianne, Emir Kamenica and Jessica Pan 2015. Gender Identity and Relative Income Within Households. *Quarterly Journal of Economics*, pp. 571-614, doi:10.1093/qje/qjv001

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

Inb Gm gith site ww ww ww Boi Soc Soc Jas Jim Bre Cur Ted Fac Che John
Secure https://www.census.gov/research/data/restricted_use_microdata.html#CRE1
Apps VRDC Gmail JMA DC AmazonSmile Expedia M&T Verified Cornell Comm Passkey ILR LDI VDI INFO7470 7470 Dev 7470 Other bookmarks

In This Section

Research Data Products

Restricted-Use Microdata

Research Data Centers

Atlanta
Boston
California - Berkeley
California - Irvine
California - Stanford
California - UCLA
California - USC
Census HQ
Central Plains
Chicago
Kansas City
Maryland
Michigan
Minnesota
Missouri
New York - Baruch
New York - Cornell

Restricted-Use Microdata

Some vital questions cannot be answered with publicly available data. The Census Bureau partners with various universities, research institutions, and agencies to form a nationwide network of secure Census Bureau Research Data Centers.

At these secure centers, qualified researchers with approved projects can perform statistical analysis on selected internal microdata from the Census Bureau and other statistical agencies.

Explore the links below to learn more about the microdata available at secure Census Bureau Research Data Centers.

- Demographic Data
  Decennial Census and surveys of individuals and households
- Economic Data
  Economic Census and surveys of businesses
- Longitudinal Employer-Household Dynamics Data
  Job-level quarterly earnings history data, data on where workers live and work, and data on firm characteristics based on combined firm and worker information

Changing Research Environment

In response to the growing literature in mathematics, computer science and statistics, which suggests that traditional statistical disclosure limitation or disclosure avoidance methods may not be strong enough to protect public-use tabular and microdata releases in the future, the Census Bureau has launched a major effort to modernize. Our current efforts focus on the 2020 Census of Population and Housing, the American Community Survey, and the 2017 Economic Census. Eventually, we will extend them to all public-use products released under the Bureau's Title 13 confidentiality pledge.

At the core these modernization efforts is a commitment to maintain the scientific integrity of the analyses performed using the public-use products that the Census Bureau releases. Historically, statistical policies requiring that published estimates include information on the margins of error associated with ensemble of methods used have excluded the effects of disclosure limitation systems. When the modernized methods are in place, the Census Bureau will issue suitability for use guidelines and margins of error that also reflect the effects of the disclosure limitation. We will publish our algorithms, and the parameter values used by those algorithms, so that researchers can use the data in a scientifically appropriate manner.

To conduct scientific analyses for which the public-use data are not suitable or sufficiently accurate, researchers may choose to seek approval to conduct a project under the auspices of the Federal Statistical Research Data Centers. Modernized disclosure limitation systems will permit such analyses to support the underlying science while remaining within the boundaries dictated by Census Bureau policies on disclosure limitation as determined by the Disclosure Review Board and the Data Stewardship Executive Committee.

This symbol indicates a link to a non-government web site. Our linking to these sites does not constitute an endorsement of any products, services or the information found on them. Once you link to another site you are subject to the policies of the new site.

The Database Reconstruction Theorem

- Powerful result from Dinur and Nissim (2003) [link]
- *Too many statistics published too accurately from a confidential database exposes the entire database with near certainty*
- How accurately is "too accurately"?
  - Cumulative noise must be of the order $\sqrt{N}$
- This theorem is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared

Database Reconstruction II

- Led quickly to “differential privacy”:
  - Dwork, McSherry, Nissim, and Smith (2006) [link]
  - Updated version (JPC 2017) [link]
  - Dwork (2006) [link]
- Leading formal privacy model
- These methods are the last soldiers standing in the search for a solution to the database reconstruction theorem

Database Reconstruction III

- "The Fundamental Law of Information Recovery"
  - Dwork and Roth, 2014 [link]
  - Dwork, undated [link]
- Includes extensions found in
  - Dwork, McSherry and Talwar (2007) [link]
  - Muthukrishnan and Nikolov (2012) [link]
  - Kasiviswanathan, Rudelson and Smith (2013) [link]
  - Dwork, Smith, Steinke, Ullman, and Vadhan (2015) [link]

Historical Note

- The U.S. Census Bureau: first organization in the world to use a formally private confidentiality protection system in production
  - OnTheMap (residential side)
- Machanavajjhala, Kifer, Abowd, Gehrke, and Vilhuber (2008) [link]

What Is a Privacy-loss Budget?

- Not a dollar budget, but works the same way
- Publishing results from confidential data has fundamental economics that cannot be swept aside
- Privacy-loss budgets constrain the aggregate risk of partial database reconstruction given all published statistics
- Worst-case limit to the inferential disclosure of any identity or item
- In differential privacy, worst case is over all possible databases with the same schema for all individuals and items

Why Use Worst-case Protection?

- "Worst case" is "equal protection under the law"
  - Protects every person in the population the same way
  - Anyone who might have been selected for the census or survey, whether in the database or not
- "Average-case" protection does not
  - Can identify who is advantaged or disadvantaged *a priori*

Respecting a Privacy-loss Budget

- All released statistics can *never* permit a database reconstruction more accurate than the budget
- Protection into the indefinite future
- For differential privacy, guarantee is over all future attackers and any database with the same schema

Current Context

- Don't current confidentiality laws require data stewards to respect a privacy-loss budget, at least implicitly?
- Unclear
- Law are silent on limitations of what can be learned about the confidential data from the released statistics (database reconstruction)
- All data publication inherently involves some inferential disclosure risk; otherwise, it is useless
  - Dwork and Naor (2008) [link]: impossibility theorem
  - Kifer and Machanavajjhala (2011) [link]: no free lunch theorem

This Is Not a New Problem

- Ratio of the circumference of a circle to its diameter is a constant
- Ancients didn't understand irrational numbers:
  - Babylonians: $\pi = 3\,{}^{1}/_{8}$
  - Egyptians: $\pi = 4 \times ({}^{8}/_{9})^{2}$
  - Israelites: $\pi = 3$ [Talmud legislated value]
  - Hindu: $\pi = \frac{62{,}832}{20{,}000} = 3.1416$
  - Euclid: no rational number is exact for this problem
  - Archimedes: sequences can approximate π with increasing accuracy
- But legal documents continued to use crude approximations
- Takes time to process abstract ideas into practical laws
- Legal guidance on inferential disclosure limitation is important
- But must be constructed sensibly

Source: Beckman, Petr "A History of Pi" (1971) [link]

Example: Randomized Response

- Randomized response is provably privacy-loss protective
- Privacy loss bounded by the maximum Bayes factor

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

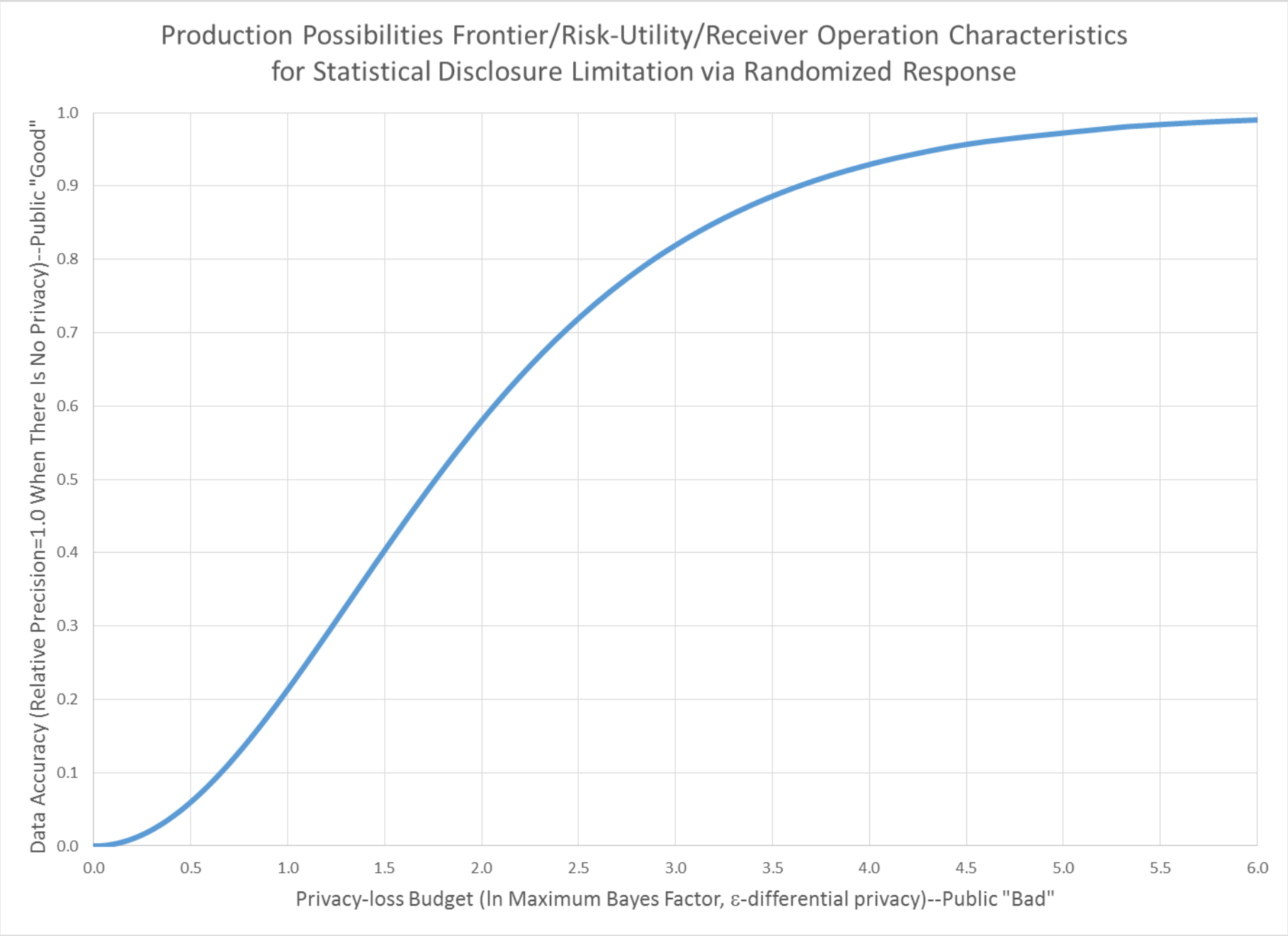
- Bound is the logarithm of the maximum Bayes factor
- If
  - Sensitive question asked with probability ½
  - And innocuous question is "yes" with probability ½
  - Then the maximum Bayes factor is 3, and ln 3 = 1.1
- The privacy-loss expenditure (ε-differential privacy) is 1.1
- Sources: Warner (1965) [link] and Greenberg, Abdel-Latif, Simmons, and Horvitz (1969) [link]. SDL uses: Fienberg and Steele (1998) [link], Du and Zhan (2003) [link] and Erlingsson, Vasyl and Korolova (2014) [link].

What Happens to Data Accuracy?

- Use relative sampling precision

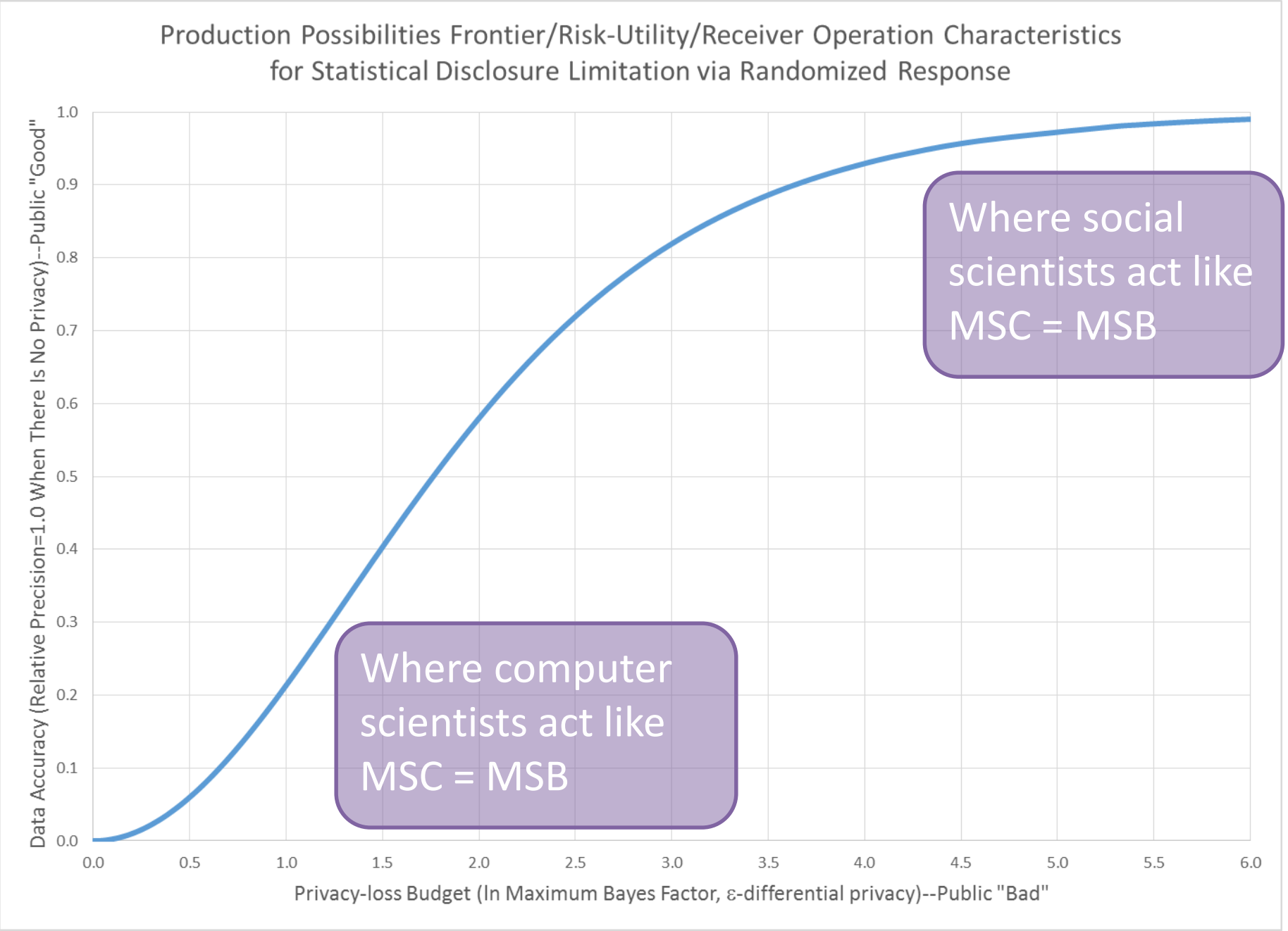
$$Rel.Precision = \frac{\{Pr[Ask\ Sensitive\ Q]\}^2 \frac{n}{\theta(1-\theta)}}{\frac{n}{\theta(1-\theta)}} = \left\{\frac{1}{2}\right\}^2 = 0.25$$

- If
  - Privacy loss is ln 3
  - Then, relative sampling precision is 25% of the most accurate estimator


Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics
for Statistical Disclosure Limitation via Randomized Response
Data Accuracy (Relative Precision=1.0 When There Is No Privacy)--Public "Good"
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
Privacy-loss Budget (ln Maximum Bayes Factor, ε-differential privacy)--Public "Bad"

Disclosure Limitation Is Technology

- The price of increasing data quality (public "good") in terms of increased privacy loss (public "bad") is the slope of the technology frontier:
  - Economics: Production Possibilities Frontier (Risk-Return in finance)
  - Forecasting models: Receiver Operating Characteristics Curve
  - Statistical Disclosure Limitation: Risk-Utility Curve (with risk on the x-axis)
- All exactly the same thing
- None able to select an optimal point


Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics
for Statistical Disclosure Limitation via Randomized Response
Data Accuracy (Relative Precision=1.0 When There Is No Privacy)--Public "Good"
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
Privacy-loss Budget (ln Maximum Bayes Factor, ε-differential privacy)--Public "Bad"
Where social scientists act like MSC = MSB
Where computer scientists act like MSC = MSB

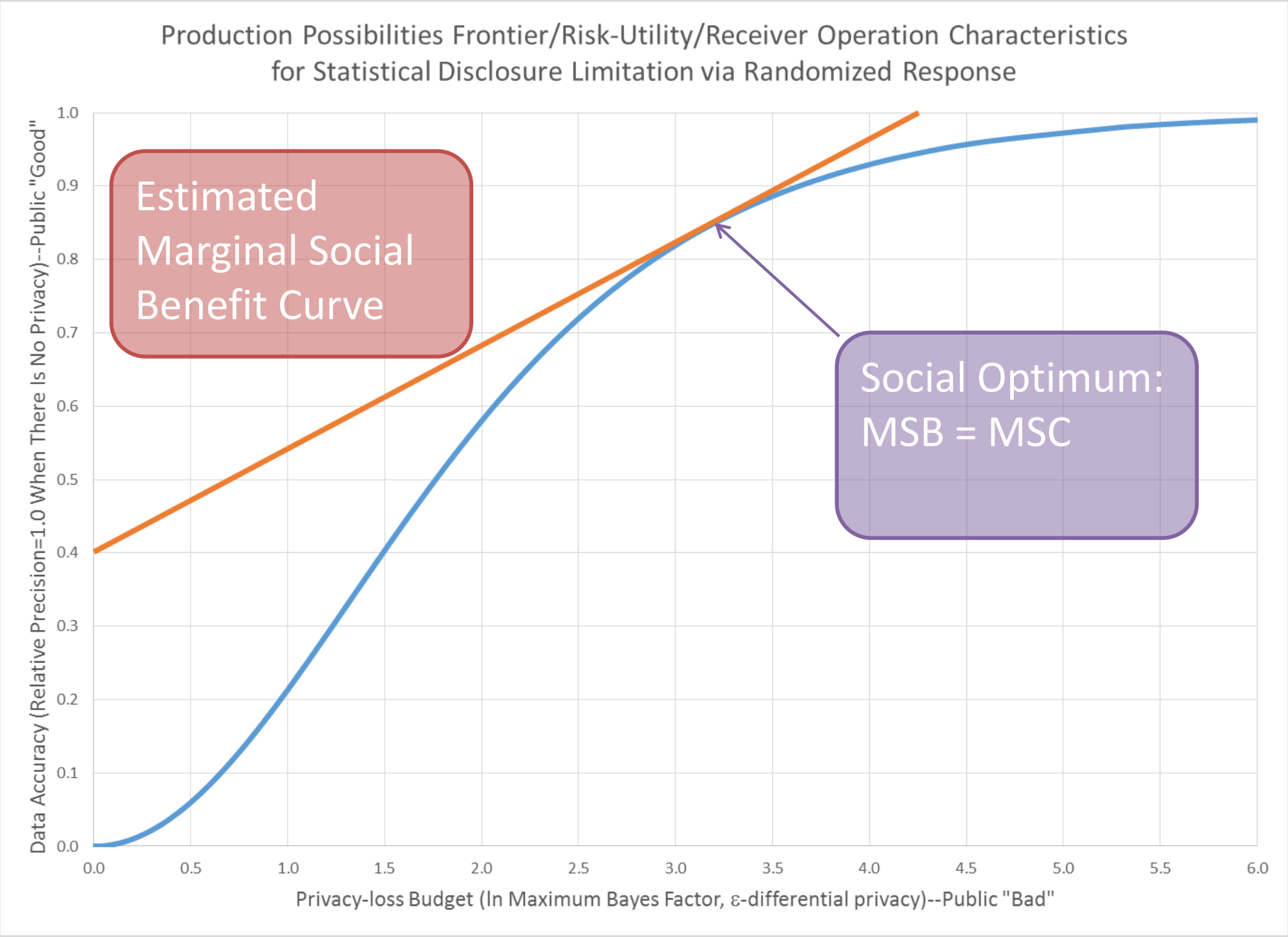
Some Examples

- Dwork (2008): "The parameter e in Definition 1 is public. The choice of ε is essentially a social question and is beyond the scope of this paper." [link, p. 3]
- Dwork (2011): "The parameter e 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." [link, p. 91]
- In OnTheMap, $\varepsilon = 8.9$, was required to produce tract-level estimates with acceptable accuracy

How to Think about the Social Choice Problem

- The marginal social benefit is the sum of all citizens' willingness-to-pay for data accuracy with increased privacy loss
- Can be estimated from survey data
- The next slide shows how

See Abowd and Schmutte (2015, revised 2017) [link]


Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics
for Statistical Disclosure Limitation via Randomized Response
Estimated Marginal Social Benefit Curve
Social Optimum: MSB = MSC
Data Accuracy (Relative Precision=1.0 When There Is No Privacy)--Public "Good"
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
Privacy-loss Budget (ln Maximum Bayes Factor, ε-differential privacy)--Public "Bad"

How to Prove That a Privacy-loss Budget Was Respected

- The privacy-loss budget captures the global disclosure risk from all publications
- Must quantify the privacy-loss expenditure of each publication
- The collection of the algorithms taken altogether must satisfy the privacy-loss budget
- Requires methods that compose

How to Prove That the Algorithms Are Resistant to All Future Attacks

- Information environment is changing much faster than before
- *It may no longer be reasonable to assert that a product is empirically safe given best-practice disclosure limitation prior to its release*
- Formal privacy models replace empirical assessment with designed protection
- Resistance to all future attacks is a property of the design

Reproducible, Private, Better Science

- American Statistical Association on p-values [link]
- Call for more nuanced use
- Data analysis conducted using privacy-preserving methods:
  - Controls the false discovery rate
  - Reduces inferential errors due to multiple comparisons
  - Examples: Erlingsson, Vasyl and Korolova (2014) [link]; Dwork et al. (2015) [link]; Apple (2016) [link]

And in the FSRDCs?

- The Census Bureau is committed to fostering reproducible science in the RDCs
  - This commitment is embodied in Data Stewardship Policy DS001, the oldest policy on record, which states that the benefit to the Census Bureau is not fully realized until the peer-reviewed publication appears
- The process that all of you go through to get results released is much more rigorous than the Bureau's internal standards
  - We are fixing that (enforcing internal standards that are the same as in FSRDCs)
  - We are using the disclosure avoidance review process as a model for a voluntary reproducibility study based on external research papers produced in the FSRDCs
- Coming soon (2020 Census will be first): end-to-end formally private analysis systems with inference validity controls

A Long Row to Hoe

- Concerted research and engineering effort needed to bring disclosure limitation into the 21st century
- Scientific integrity requires that we tackle this challenge
- First step is experimentation with the technologies known to work:
  - Synthetic data with validation using formally private synthesizers
  - Privacy-preserving data analysis via pre-specified query systems
- This is a partnership with the world-wide research community

References (in order of appearance)

- Abowd, John M. 2017. How Will Statistical Agencies Operate When All Data are Private? *Journal of Privacy and Confidentiality*. [download]
- Abowd, John M. 2016. Why Statistical Agencies Need to Take Privacy-loss Budgets Seriously, and What It Means When They Do. Federal Committee on Statistical Policy, Policy Conference. [download]
- Wasserstein, Ron L. and Nicole A. Lazar. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician, 70*, 129-133. DOI: 10.1080/00031305.2016.1154108.
- Bertrand, Marianne, Emir Kamenica and Jessica Pan 2015. Gender Identity and Relative Income Within Households. Quarterly Journal of Economics, pp. 571-614, doi:10.1093/qje/qjv001.
- Dinur, Irit and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*(PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. in Halevi, S. & Rabin, T. *(Eds.)* Calibrating Noise to Sensitivity in Private Data Analysis *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg*, 265-284, DOI: 10.1007/11681878_14.
- Dwork, Cynthia. 2006. Differential Privacy, *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006), Springer Verlag, 4052*, 1-12, ISBN: 3-540-35907-9.
- Dwork, Cynthia and Aaron Roth. 2014. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science. Vol. 9, Nos. 3–4. 211–407, DOI: 10.1561/0400000042. [download]
- Dwork, Cynthia. Undated. The State of the Art. Slide presentation. [download]
- Dwork, Cynthia, Frank McSherry and Kunal Talwar. 2007. The price of privacy and the limits of LP decoding. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*(STOC '07). ACM, New York, NY, USA, 85-94. DOI:10.1145/1250790.1250804.
- Muthukrishnan, S. and Aleksandar Nikolov. 2012. Optimal Private Halfspace Counting via Discrepancy, *CoRR (Computing Research Repository), abs/1203.5453.* [download]
- Kasiviswanathan, Shiva Prasad, Mark Rudelson and Adam Smith. 2012. The Power of Linear Reconstruction Attacks, *CoRR (Computing Research Repository),* abs/1210.2381.
- Dwork, Cynthia, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. 2015. "Robust Traceability from Trace Amounts." In IEEE Symposium on Foundations of Computer Science (FOCS 2015). Berkeley, California, 10/18-20/2015.
- Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd , Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.
- Dwork, Cynthia and Moni Naor. 2010. On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy, *Journal of Privacy and Confidentiality*: Vol. 2: Iss. 1, Article 8, pp. 93-107. [download]
- Kifer, Daniel and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (SIGMOD '11). ACM, New York, NY, USA, 193-204. DOI:10.1145/1989323.1989345.
- Beckman, Petr. 1971. *A History of Pi*. Barnes and Noble, New York, NY. ISBN:0-88029-418-3.
- Warner, Stanley. L. 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias *Journal of the American Statistical Association, 60*, 63-69.
- Greenberg, Bernard G., Abdel-Latif Abul-Ela, Walt R. Simmons, and Daniel G. Horvitz. 1969. The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association, 64*, 520-539.
- Fienberg, Stephen E. and Russel J. Steele. 1998. *Journal of Official Statistics,* Vol. 14, No. 4 (December): 485-502.
- Du, Wenliang and Zhijun Zhan. 2003. Using randomized response techniques for privacy-preserving data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '03). ACM, New York, NY, USA, 505-510. DOI:10.1145/956750.956810.
- Erlingsson, Úlfar, Vasyl Pihur and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (CCS '14). ACM, New York, NY, USA, 1054-1067. DOI:10.1145/2660267.2660348.
- Dwork, Cynthia. 2008. Differential Privacy: A Survey of Results. In Agrawal, M., Du, D.; Duan, Z. and Li, A. *(Eds.). Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings, Springer Berlin Heidelberg*, 1-19. [download]
- Dwork, Cynthia. 2011. A firm foundation for private data analysis. *Communications ACM* 54, 1 (January): 86-95. DOI:10.1145/1866739.1866758.
- Abowd, John M. and Ian M. Schmutte. 2017. Revisiting the economics of privacy: Population statistics and confidentiality protection as public goods. Labor Dynamics Institute, Cornell University, Labor Dynamics Institute, Cornell University. [download]
- Dwork, Cynthia, Feldman V., Hardt M., Pitassi T., Reingold O., Roth A. 2015. The reusable holdout: Preserving validity in adaptive data analysis. Science, 349, 6248 (August 7):636-8. DOI:10.1126/science.aaa9375.
- Apple, Inc. 2016. Apple previews iOS 10, the biggest iOS release ever. Press Release (June 13). URL=http://www.apple.com/newsroom/2016/06/apple-previews-ios-10-biggest-ios-release-ever.html.
- Haney, Samuel, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. 2017. Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics, SIGMOD, DOI: http://dx.doi.org/10.1145/3035918.3035940

Thank you

john.maron.abowd@census.gov

Working Current Examples

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

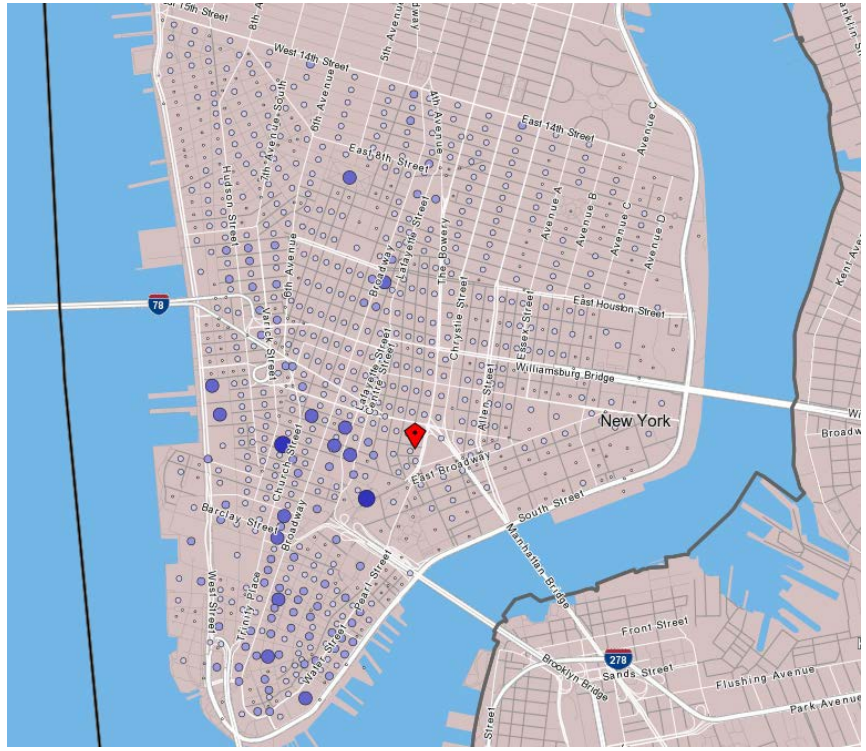
Example 1: OnTheMap Employer Data

Background/Motivation

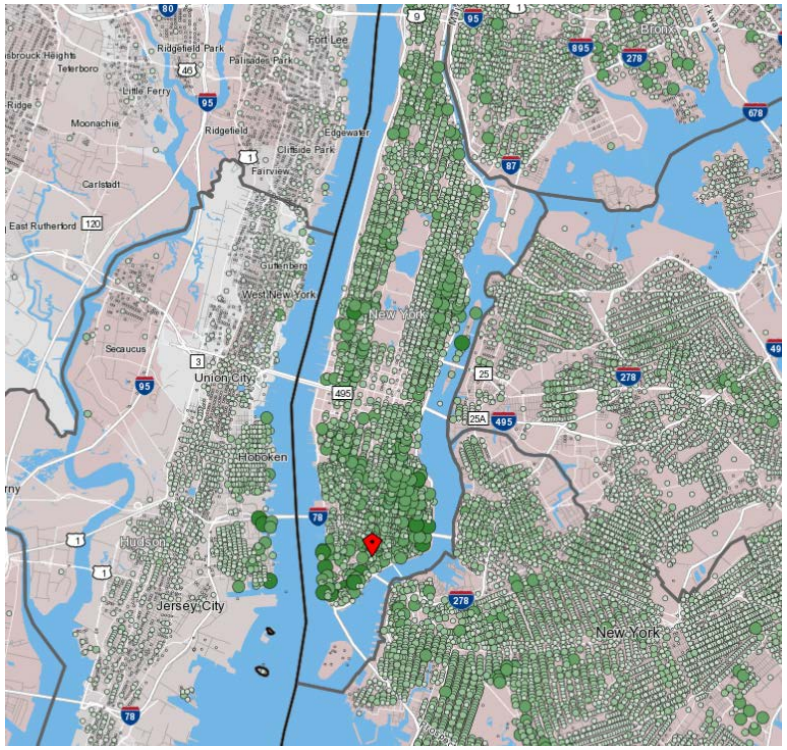
- Currently, the Census Bureau releases the LEHD Origin-Destination Employment Statistics (LODES).
- In protecting confidentiality of workers and firms, LODES utilizes:
  - Permanent multiplicative noise distortion factors at the employer and establishment level (plus synthetic data methods for small 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

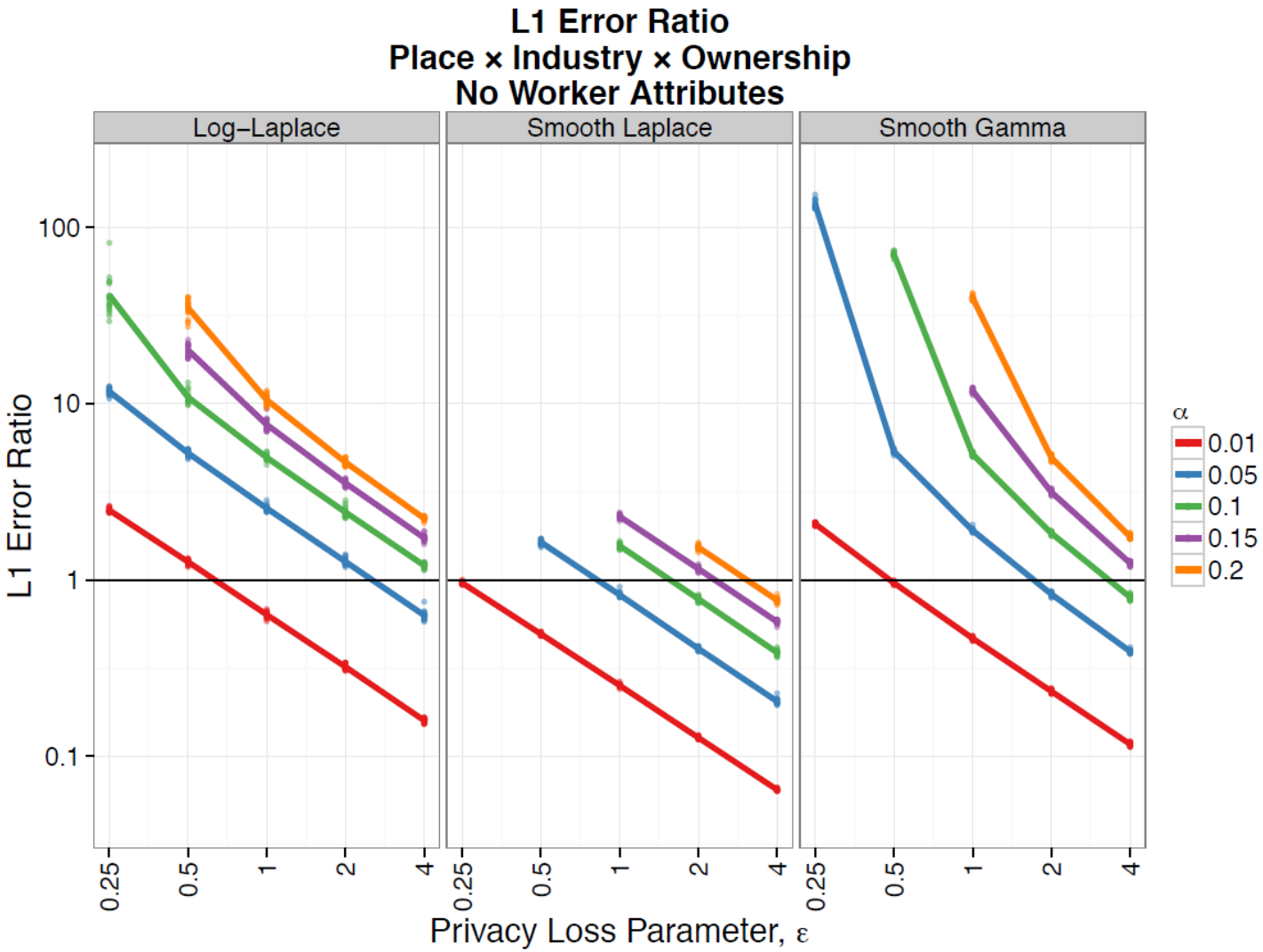


Available at http://onthemap.ces.census.gov/

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**

Source: Haney et al. SIGMOD 2017.



Source: Haney et al. 2017

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)

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