Why the Economics Profession Cannot Cede the Discussion of Privacy Protection to Computer Scientists

John M. Abowd
U.S. Census Bureau and Cornell University
Joint work with Ian M. Schmutte, William N. Sexton and Lars Vilhuber
Allied Social Science Associations Annual Meetings
The Future of Economic Research under Rising Risks and Costs of Information Disclosure
Saturday 2:30-4:30, January 5, 2019

United States™ Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

The views expressed in this presentation are those of the authors and not those of the U.S. Census Bureau or other sponsors.

You Must Remember This [with apologies to Herman Hupfeld]

- Modern statistical disclosure limitation is only 12 years old, but it is already better than what it replaces
  - Scientific inferences from the public data are *guaranteed* consistent with the confidential data
  - *Nothing* about the process is *secret* except the “password”
  - And *they work*
- Traditional statistical disclosure limitation is broken
  - Scientific inferences from the public data have *unknown validity*
  - *Information needed* to make those inferences valid is *secret*
  - And *they no longer work*
- It took cryptography decades to get to this stage (pre-RSA)

Three Lessons from Cryptography

1. Too many statistics, published too accurately, expose the confidential database with near certainty (database reconstruction)
2. Add noise to every statistic, calibrated to control the worst-case global disclosure risk, called a privacy-loss budget (formal privacy)
3. Transparency can't be the harm: Kerckhoffs's principle applied to data privacy says that the protection should be provable and secure even when every aspect of the algorithm and all of its parameters are public, only the actual random numbers used must be kept secret

The database reconstruction theorem is the death knell for traditional data publication systems from confidential sources.

And One Giant Silver Lining

- Formal privacy methods like differential privacy provide technologies that quantify the relationship between accuracy (in multiple dimensions) and privacy-loss
- When you use formal methods, the scientific inferences are valid, provided the analysis incorporates the noise injection from the confidentiality protection
- Traditional SDL doesn’t, and can’t, do this—it is inherently scientifically dishonest

Database Reconstruction and Re-identification Attacks

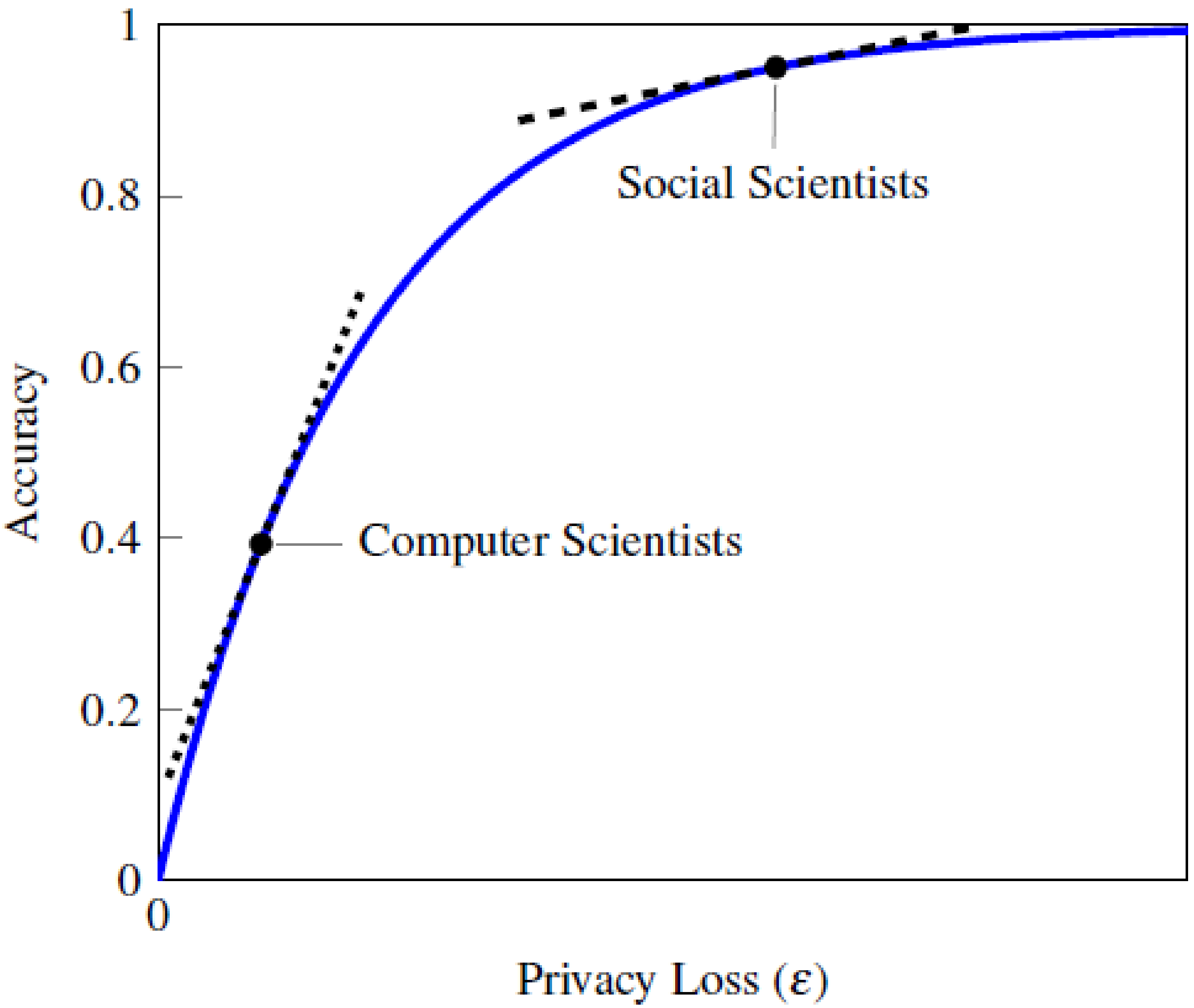
The Flaw in Traditional SDL

- Assumes that disclosure avoidance modifications made for two products from the same confidential data are compatible
- For the 2010 Census of Population, 7.8 billion linearly independent statistical summaries were published (including zeros) from 308,745,538 people
  - Geographic detail: block, tract, county
  - Other detail: age in years, OMB race/ethnicity, sex, relationship to householder, in Summary File 2: detailed race data
- A 10% Public-use micro-data sample was also released
  - Geographic detail: Public-use micro-area (PUMA, 100,000 population minimum)
  - Other detail as in tabular data with some modifications
- The tabular summaries can be used to produce a 100% micro-data file
  - Geographic detail: block, tract, county
  - Other detail: exactly as in the tabular data
- Database reconstruction was used to create the 100% micro-data file

Internal Experiments Using the 2010 Census

- Confirm that the micro-data from the confidential 2010 Hundred-percent Detail File (HDF) can be accurately reconstructed from PL94 + balance of SF1
  - The block geocode is always correct, as is voting age
  - 50% of reconstructed records match both the HDF (swapped) and Census Edited File (CEF, unswapped) exactly
  - More than 70% match exactly if age +/- one year is used
  - More than 90% match if one match error in age, sex, race or ethnicity is allowed
  - More than 50% of the confidential records (CEF) are population uniques on block, age (in years), sex, race (OMB) and ethnicity
  - More than 35% of the reconstructed records are population uniques on the same variables
- While there is a reconstruction vulnerability, the risk of re-identification is apparently still relatively small
  - A much larger percentage were successfully re-identified than in any previous experiments
  - The major impediment to re-identification is poor quality geocodes on the Census Bureau copies of commercial databases
- Experiments are at the person level, not household
- Census Bureau declared reconstruction of Title 13-sensitive data an issue, no longer a risk
- Reason for the adoption of differential privacy for the 2018 End-to-End Census Test and 2020 Census
- Quantitative details are being withheld to permit external peer-review before they are released

Managing the Tradeoff between Accuracy and Privacy Loss


1
0.8
0.6
0.4
0.2
0
Accuracy
0
Privacy Loss (ε)
Social Scientists
Computer Scientists

Basic Principles

- Based on recent economics (2019, *American Economic Review*) https://digitalcommons.ilr.cornell.edu/ldi/48/ or https://arxiv.org/abs/1808.06303
- The marginal social benefit is the sum of all persons' willingness-to-pay for data accuracy with increased privacy loss
- The marginal rate of transformation is the slope of the privacy-loss v. accuracy graphs we have been examining
- This is exactly the same problem being addressed by Google in RAPPOR or PROCHLO, Apple in iOS 11, and Microsoft in Windows 10 telemetry

Moving Forward

- Proposal for vetting traditional and formally private SDL
- "Gold standard" is inference on the original confidential data
- Replication protocol
  - Establish public-use inputs
  - Verify computations produce original published results
- Validation protocol
  - Establish original confidential inputs to public-use products
  - Validate computations from replication protocol on confidential inputs
- Comparison protocol
  - Use the replication protocol on proposed new public-use products
  - Use the validation protocol for proposed new products
  - Compare validated original products to validated new products

Thank you.

John.Maron.Abowd@census.gov

Four Examples from Abowd & Schmutte (2019)

- Legislative redistricting
- Economic censuses and national accounts
- Tax data and tax simulations
- General purpose public-use micro-data

Legislative Redistricting

- In the redistricting application, the fitness-for-use is based on
  - Supreme Court one-person one-vote decision (All legislative districts must have approximately equal populations; there is judicially approved variation)
  - *Is statistical disclosure limitation a “statistical method” (permitted by Utah v. Evans) or “sampling” (prohibited by the Census Act, confirmed in Commerce v. House of Representatives)?*
  - Voting Rights Act, Section 2: requires majority-minority districts at all levels, when certain criteria are met
- The privacy interest is based on
  - Title 13 requirement not to publish exact identifying information
  - The public policy implications of uses of race, ethnicity and citizenship tabulations at detailed geography

Economic Censuses and National Accounts

- The major client for the detailed tabulations from economic censuses is the producer of national accounts
- In most countries these activities are consolidated in a single agency
- Fitness-for-use: accuracy of the national accounts
- Privacy considerations: sensitivity of detailed industry and product data, which may have been supplied by only a few firms
- Detailed tables can be produced using formal privacy, with far less suppression bias than in current methods
- But, its an inefficient use of the global privacy-loss budget when the accounts are published at much more aggregated levels
- Optimize the accuracy v. privacy loss by sharing confidential data (as permitted under CIPSEA) and applying formal privacy at publication level

Tax Data and Tax Simulations

- Simulating the effects of tax policy changes is an important use of tax micro-data
- Traditional disclosure limitation methods aggravate these simulations by smoothing over important kinks and breaking audit consistency
- Fitness-for-use: quality of the simulated tax policy effects
- Privacy: sensitivity of the income tax returns
- Optimize the accuracy v. privacy loss by doing the simulations inside the IRS firewall and applying formal privacy protection to outputs

General Purpose Public-use Micro-data

- Hierarchy of users
  - Educational
  - Commercial
  - Scientific
- Fitness-for-use: valid scientific inferences on arbitrary hypotheses estimable within the design of the confidential data product
- Privacy: database reconstruction-abetted re-identification attacks make every variable a potential identifier, especially in combination
- Traditional SDL fails the fitness-for-use
- Formal privacy guarantees the fitness-for-use for hypotheses in the set supported by its query workload (serves educational and commercial uses very well)
- For other hypotheses, supervised use (perhaps via a validation server) maintains fitness for use
- Same model as used by IPUMS https://international.ipums.org/international/irde.shtml
- We need to build these cooperatively

More Background on the 2020 Disclosure Avoidance System

- September 14, 2017 CSAC (overall design) https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf?#
- August, 2018 KDD'18 (top-down v. block-by-block) https://digitalcommons.ilr.cornell.edu/ldi/49/
- October, 2018 WPES (implementation issues) https://arxiv.org/abs/1809.02201
- October, 2018 *ACMQueue* (understanding database reconstruction) https://digitalcommons.ilr.cornell.edu/ldi/50/ or https://queue.acm.org/detail.cfm?id=3295691
- December 6, 2010 CSAC (detailed discussion of algorithms and choices) https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#

Selected References

- 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.
- 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.
- 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. Available at: http://repository.cmu.edu/jpc/vol2/iss1/8.
- 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.
- Abowd, John M. and Ian M. Schmutte. Forthcoming. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. American Economic Review, at https://arxiv.org/abs/1808.06303
- 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.
- 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.
- Ding, Bolin, Janardhan Kulkarni, and Sergey Yekhanin 2017. Collecting Telemetry Data Privately, NIPS 2017.
- Bittau , Andrea, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Usharsee Kode, Julien Tinnes, and Bernhard Seefeld 2017. Prochlo: Strong Privacy for Analytics in the Crowd, https://arxiv.org/abs/1710.00901.