Preparing for the 2020 Census: Disclosure Avoidance

John M. Abowd
Associate Director for Research and Methodology, and Chief Scientist
U.S. Census Bureau
American Association of Geographers Annual Meeting
April 4, 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 talk are my own and not those of the U.S. Census Bureau.

Confidentiality Protection in the 21st Century

- The publications from the 2020 Census will be the best protected information product the Census Bureau can produce
- Each product will be protected in a manner that ensures its fitness for use
- We have determined that detailed publications from universe databases like a population census are vulnerable to reconstruction-abetted re-identification attacks
- The formal privacy system we are using, called differential privacy, effectively protects against this risk without compromising the fitness for use as much as hardened traditional methods would

Establishing the Risks: What We Did

- Database reconstruction for all 308,745,538 people in 2010 Census
- Link reconstructed records to commercial databases: acquire PII
- Successful linkage to commercial data: putative re-identification
- Compare putative re-identifications to confidential data
- Successful linkage to confidential data: confirmed re-identification
- Harm: attacker can learn self-response race and ethnicity

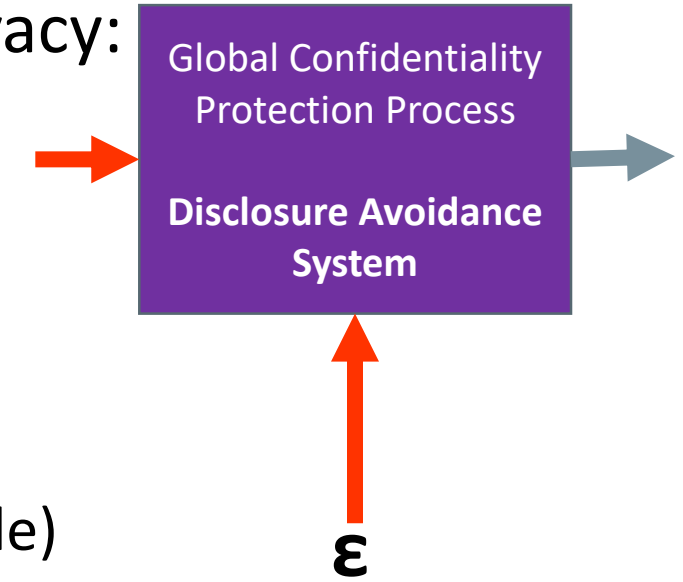
Confirming the Risks: What We Found

- Census block and voting-age correctly reconstructed in all 6,207,027 inhabited blocks
- Block, sex, age, race, ethnicity reconstructed
  - Exactly: 46% of population (142 million of 308,745,538 records in CEF)
  - Allowing age +/- one year: 71% of population (219 million of 308,745,538)
- Block, sex, age linked to commercial data to acquire PII
  - Putative re-identifications: 45% of population (138 million of 308,745,538)
- Name, block, sex, age, race, ethnicity compared to confidential CEF
  - Confirmed re-identifications: 38% of putative (52 million; 17% of population)
- For the confirmed re-identifications, race and ethnicity are learned exactly, not statistically

Differential Privacy

The Disclosure Avoidance System Relies on Injecting Noise with Formal Privacy Rules

- Advantages of noise injection using differential privacy:
  - Privacy guarantees are *closed under composition*
  - Privacy guarantees are *robust to post-processing*
  - Privacy guarantees are *future-proof*
  - Privacy guarantees are *provable and tunable*
  - Privacy guarantees are *public and explainable*
  - Protects against database reconstruction attacks (tunable)

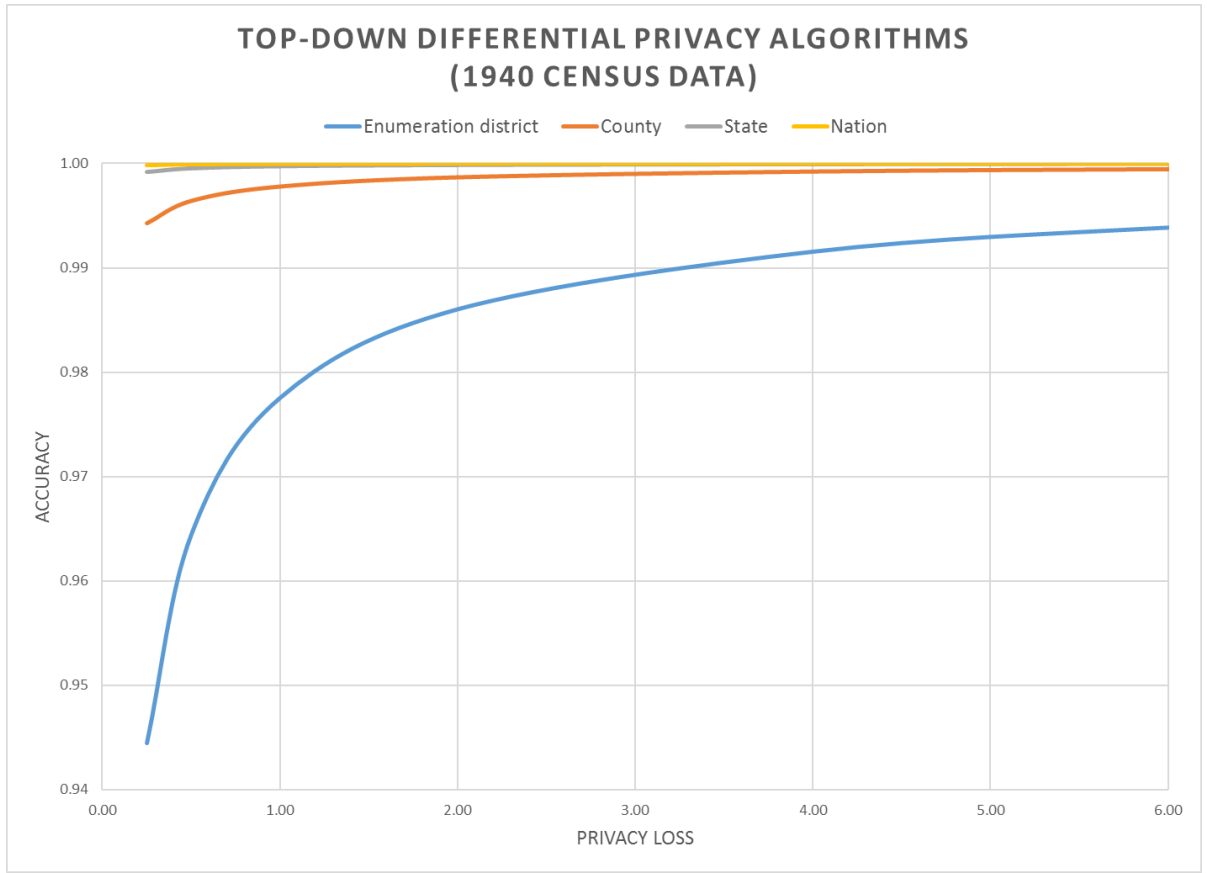


- Disadvantages:
  - Entire country must be processed at once for best accuracy
  - Every use of the private data must be tallied in the *privacy-loss budget*

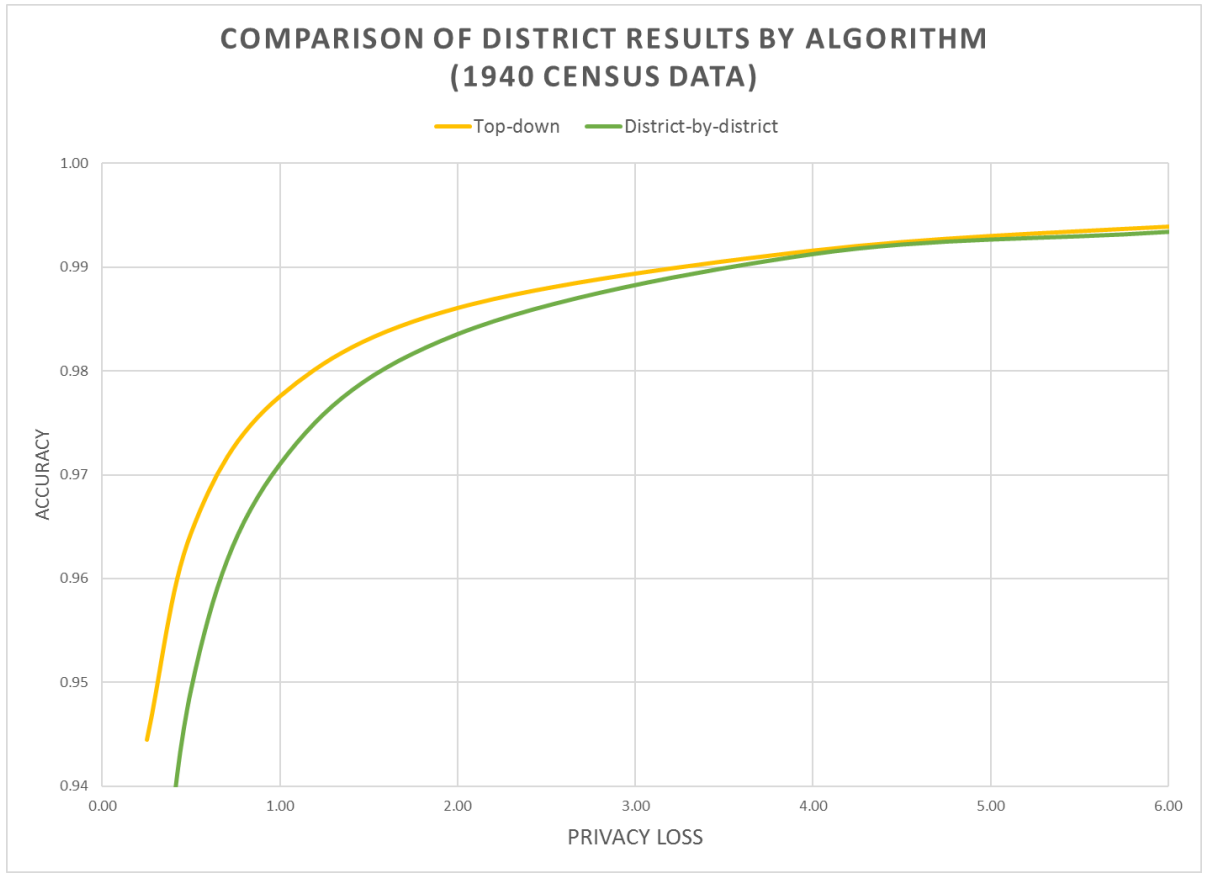
Algorithm Comparison Using Public 1940 Census Data

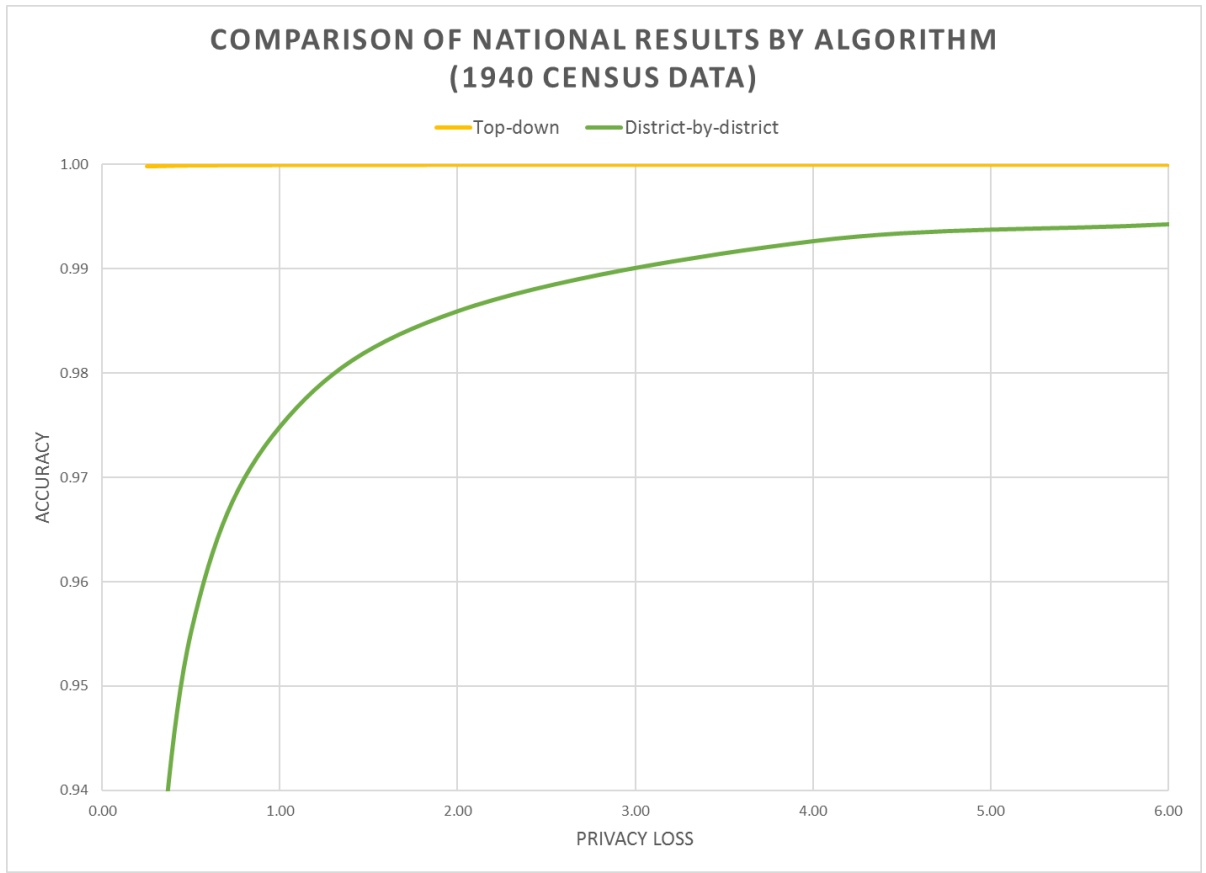
Two Candidate Algorithms

- Block-by-block (also called bottom-up)
  - DP applied to all tables at the most detailed geographic level (blocks)
  - All aggregations built from those tables
- Top-down
  - DP measurements taken at all levels of the geographic hierarchy
  - Large-scale optimization problem solved to allocate microdata records to solution tables respecting invariants, table consistency, non-negativity, and integer constraints
- In these tests, all invariants were imposed at the enumeration-district level (similar to the modern definition of block groups)


TOP-DOWN DIFFERENTIAL PRIVACY ALGORITHMS
(1940 CENSUS DATA)
Enumeration district
County
State
Nation
1.00
0.99
0.98
0.97
0.96
0.95
0.94
ACCURACY
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS


DISTRICT-BY-DISTRICT DIFFERENTIAL PRIVACY ALGORITHMS
(1940 CENSUS DATA)
Enumeration district
County
State
Nation
1.00
0.99
0.98
0.97
0.96
0.95
0.94
ACCURACY
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS

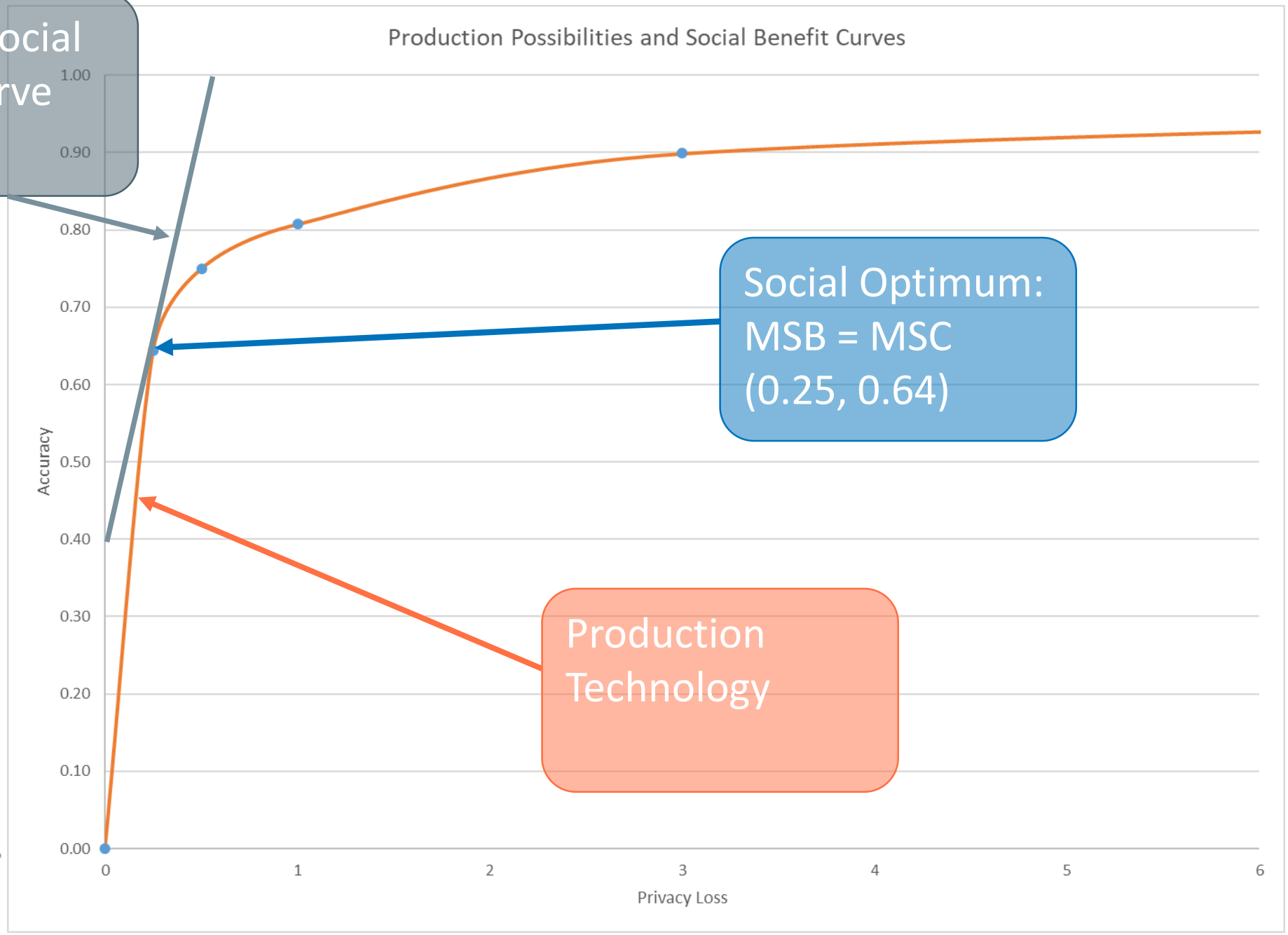

COMPARISON OF DISTRICT RESULTS BY ALGORITHM
(1940 CENSUS DATA)
Top-down
District-by-district
1.00
0.99
0.98
0.97
0.96
0.95
0.94
ACCURACY
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS


COMPARISON OF NATIONAL RESULTS BY ALGORITHM
(1940 CENSUS DATA)
Top-down
District-by-district
1.00
0.99
0.98
0.97
0.96
0.95
0.94
ACCURACY
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS

Managing a Global Privacy-loss Budget

- There are three generic uses of the global privacy-loss budget
  - Person-level queries
    - Bulk of PL94-171 and Citizen Voting-Age Population (CVAP) tables
    - Many Demographic and Housing Characteristics (DHC) tables
    - Some tables using detailed race and ethnicity, AIAN
  - Household-level queries
    - One PL94-171 table, no CVAP tables
    - Many DHC tables
    - Most tables in detailed race, ethnicity and AIAN products
  - Household-person queries
    - None in PL94-171 nor CVAP
    - Balance of tables in DHC
- Public-use microdata would be developed from these queries, so there is no additional privacy-loss

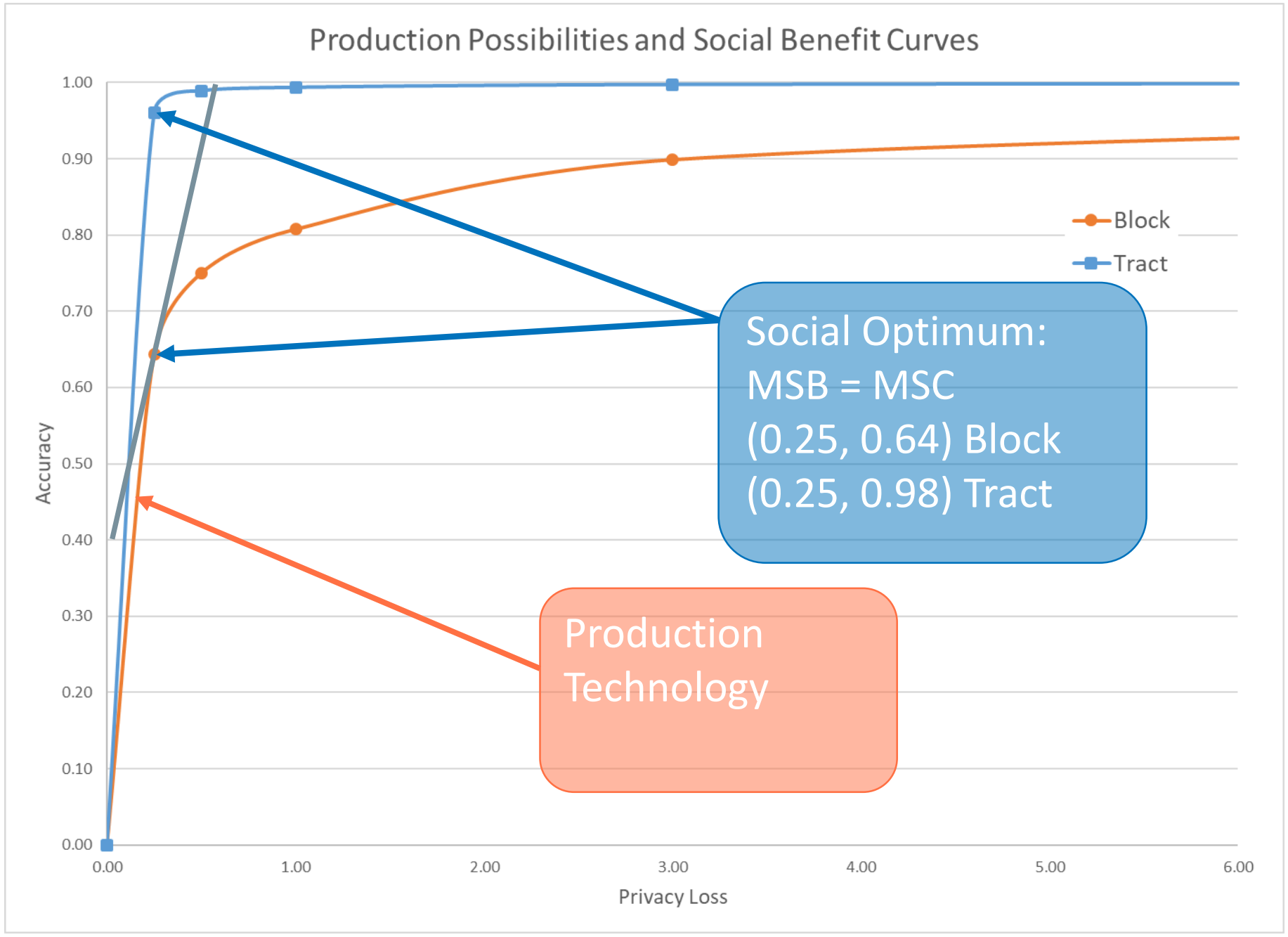
Marginal Social Benefit Curve



Social Optimum:
MSB = MSC
(0.25, 0.64)

Production Technology

United States Census Bureau



United States Census Bureau

The Importance of Formal Privacy

- Block-level summary data from the decennial census have a long history, an important and valid use case, and can be delivered with the current formal privacy system, as demonstrated in the 2018 End-to-End Census test
- Abandoning formal privacy for any part of the 2020 Census publications exposes the entire set of publications, including the block-level tables, to the same reconstruction-abetted re-identification attack strategy to which the 2010 Census was vulnerable
- The current environment is equivalent to exposing a major cybersecurity vulnerability: you can’t patch one part and leave other parts exposed—you have to fix the whole system

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, 2018 CSAC (block-level implementation) https://www.census.gov/about/cac/sac/meetings/2018-12-meeting.html
- March 28, 2018 CSAC (managing the privacy-loss budget) https://www.census.gov/about/cac/sac/meetings/2019-03-meeting.html

Thank you.

John.Maron.Abowd@census.gov

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. 2019. 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.

Backup slides

Technical Challenges

- Hierarchical and table consistency
- Invariants (main algorithmic challenge in combination with hierarchy)
- Asymptotic consistency as ε (privacy loss) gets large
- Existence proofs for solutions
- Workload optimization, especially for join tables (persons x households)
- Many implementation issues, not discussed here

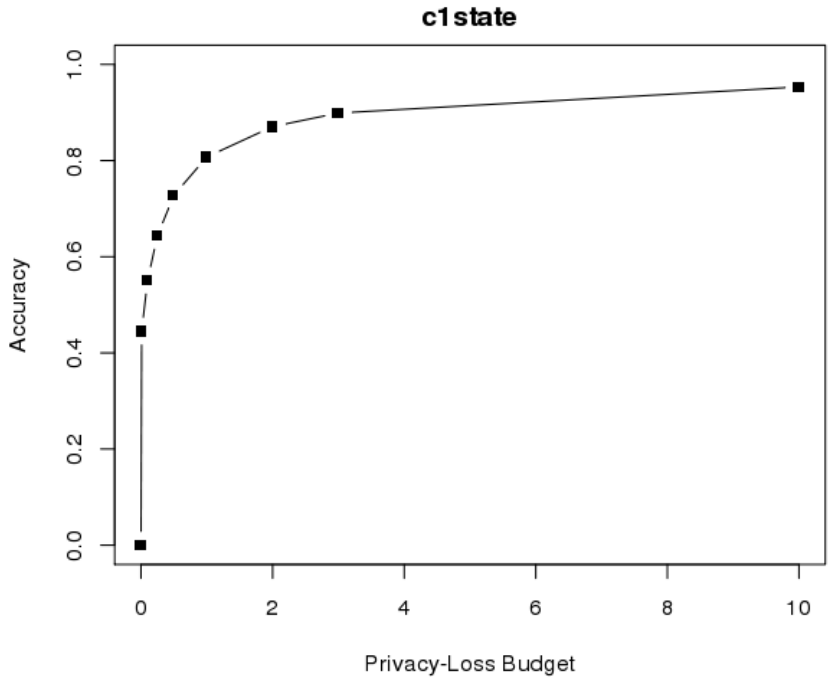
Analyses Supporting the 2018 End-to-End Census Test

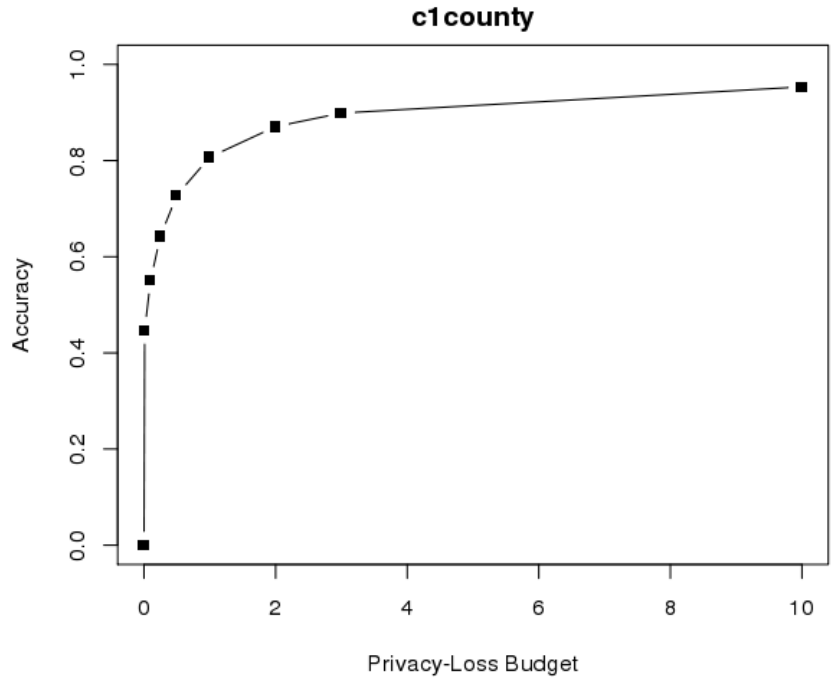
Implementation Decisions for 2018 End-to-End Census Test

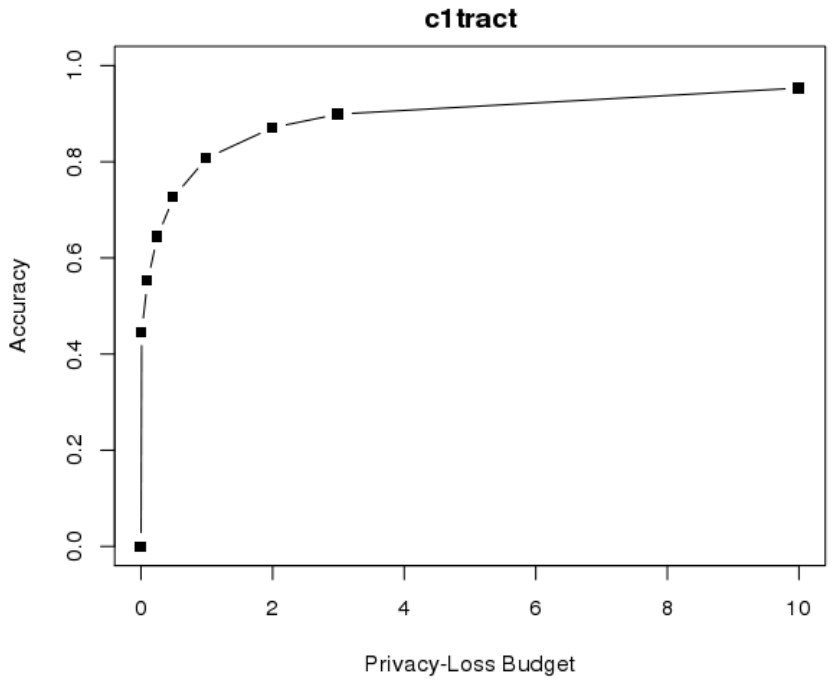
- Population invariant at the county level (Providence, RI is the only county in the test)
- No voting-age invariant at any level
- Number of housing units invariant down to the block (design constraint due to operation of LUCA and address canvassing)
- Number of occupied housing units invariant (could not be relaxed in time to meet E2E production deadlines)
- Number and type of group quarters invariant down to the block (same design constraint as number of housing units)
- Global privacy-loss budget (ε) 0.25
- Allocation to PL94-171 100% (no other tables being released)

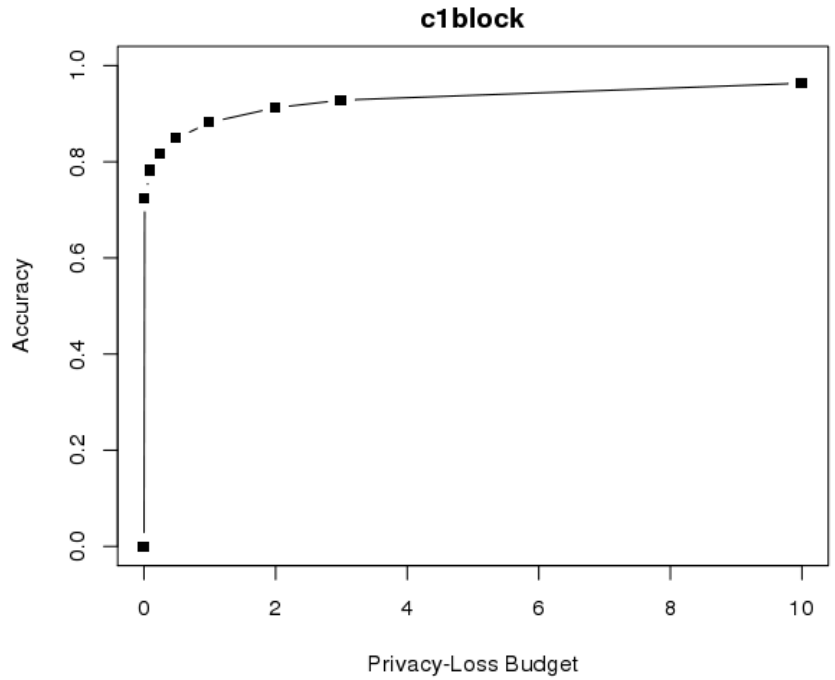
Accuracy v. Privacy Loss for Rhode Island (2010 Census) using the 2018 E2E Test Disclosure Avoidance System

PL94-171 redistricting data at the block level





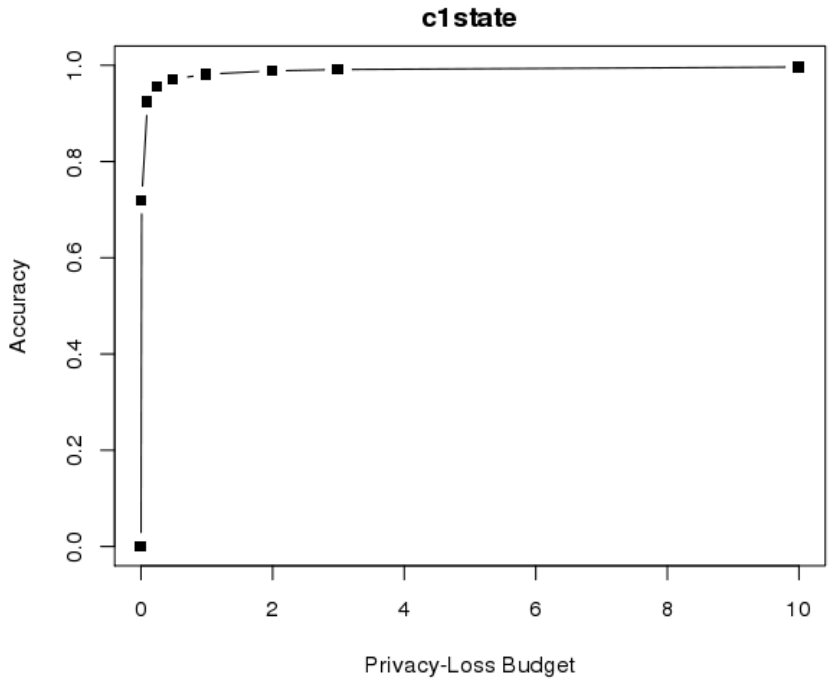


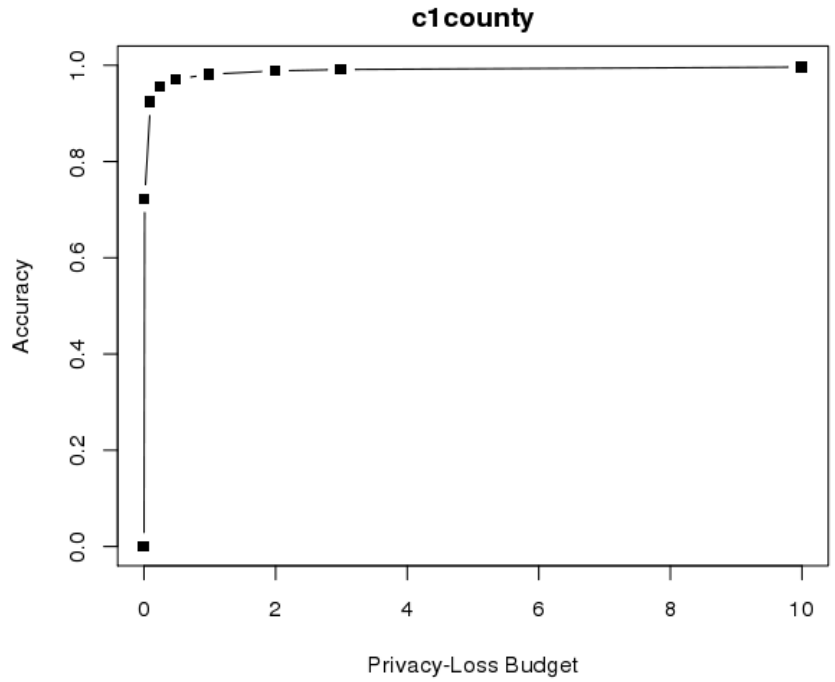


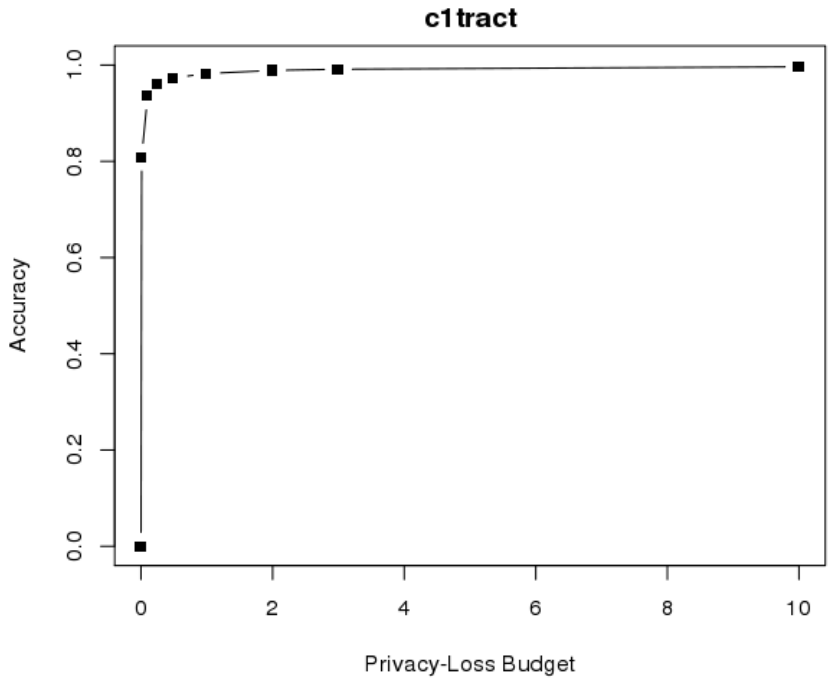
United States Census Bureau

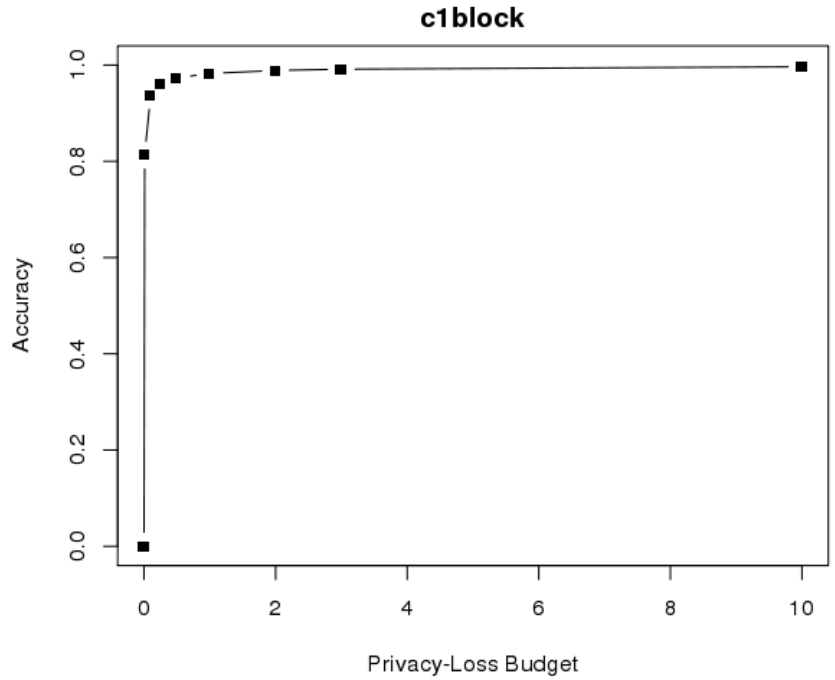
Accuracy v. Privacy Loss for Rhode Island (2010 Census) using the 2018 E2E Test Disclosure Avoidance System

PL94-171 redistricting data at the tract level





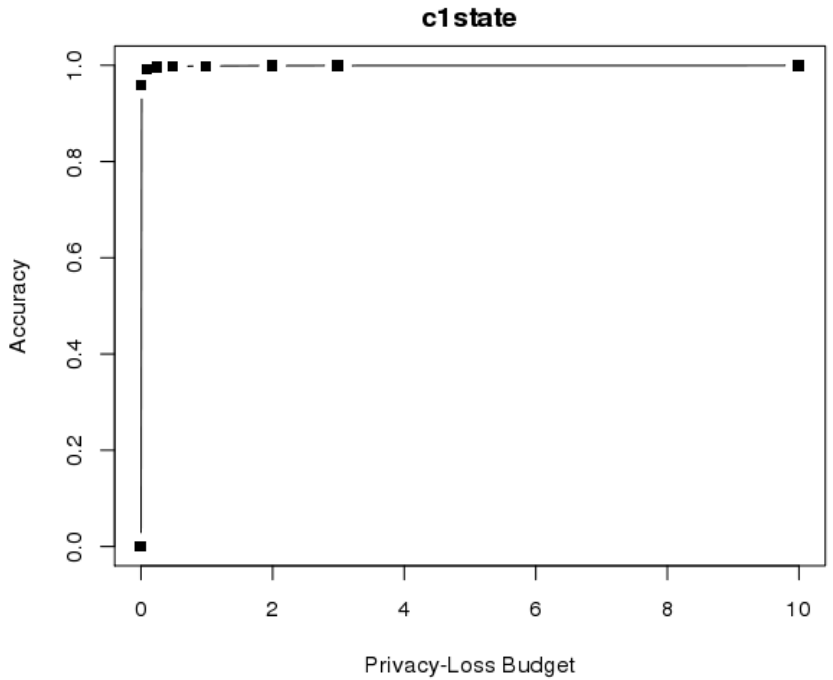


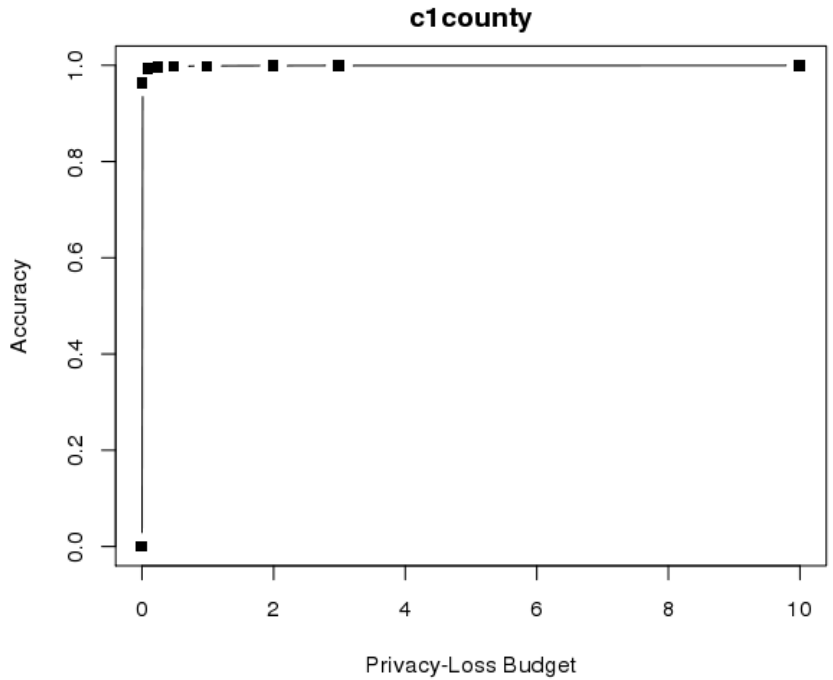


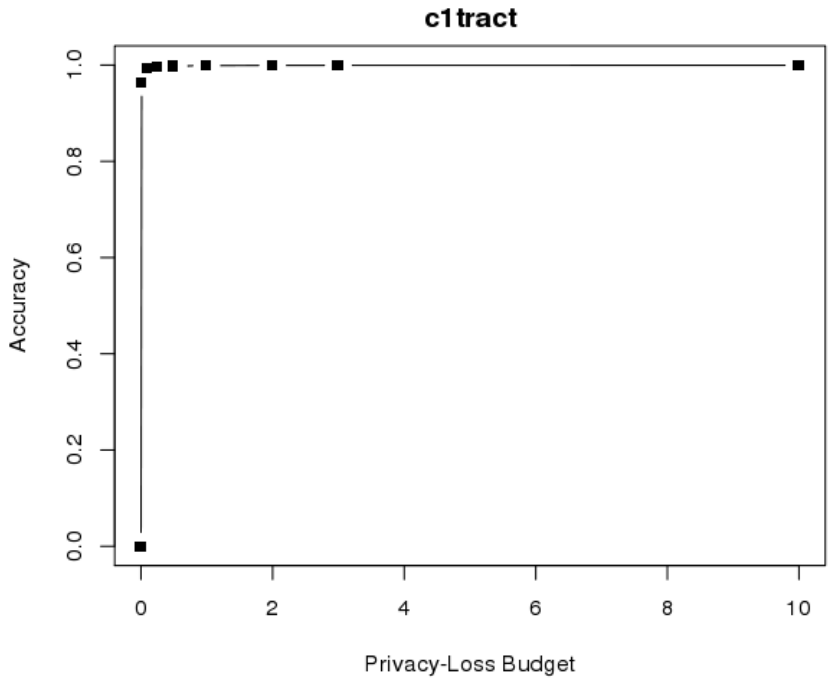
United States Census Bureau

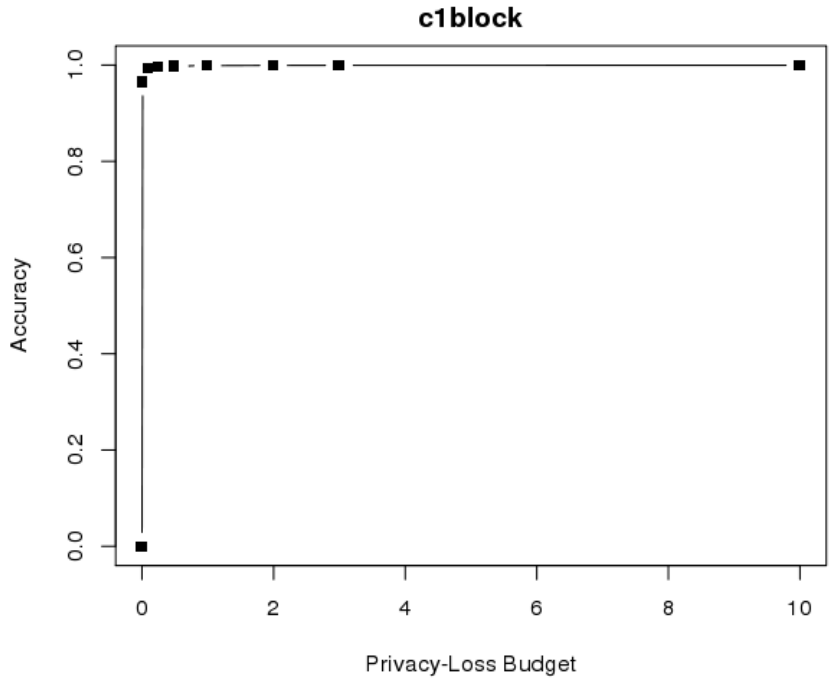
Accuracy v. Privacy Loss for Rhode Island (2010 Census) using the 2018 E2E Test Disclosure Avoidance System

PL94-171 redistricting data at the county level





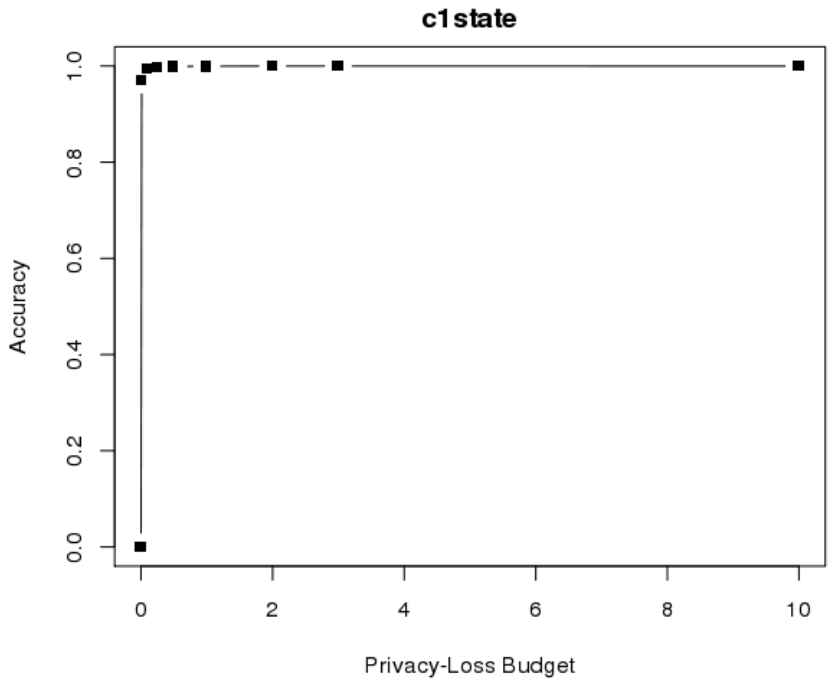


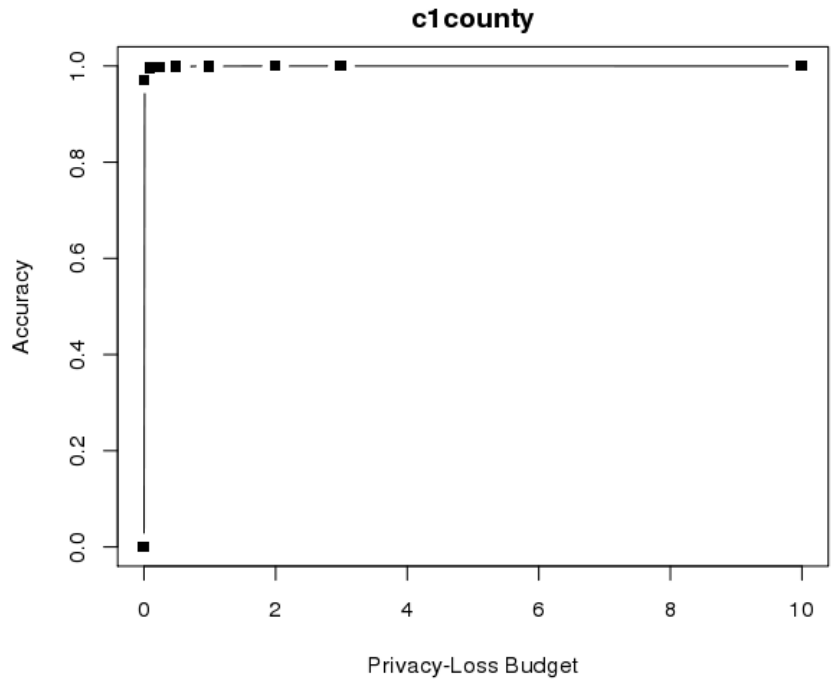


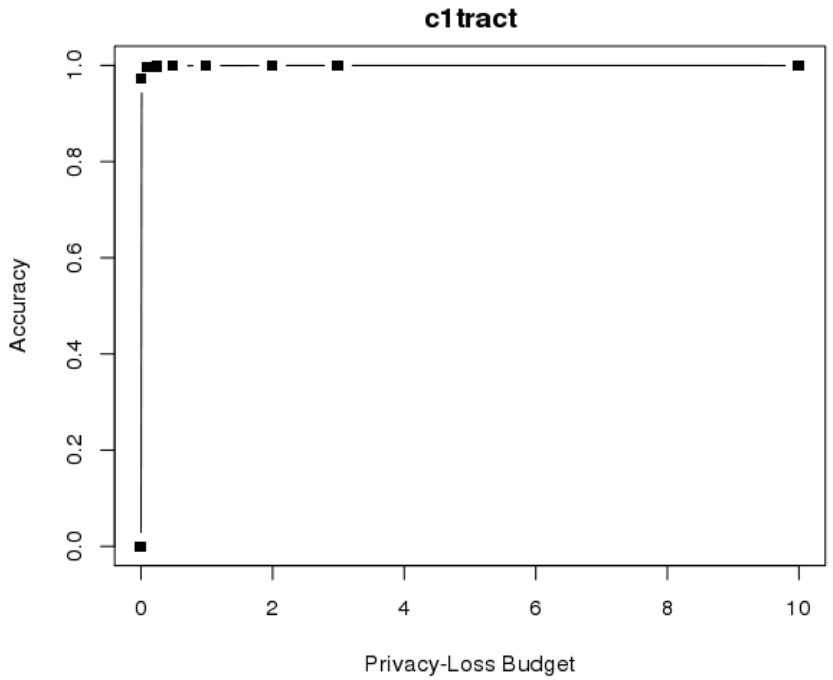
United States Census Bureau

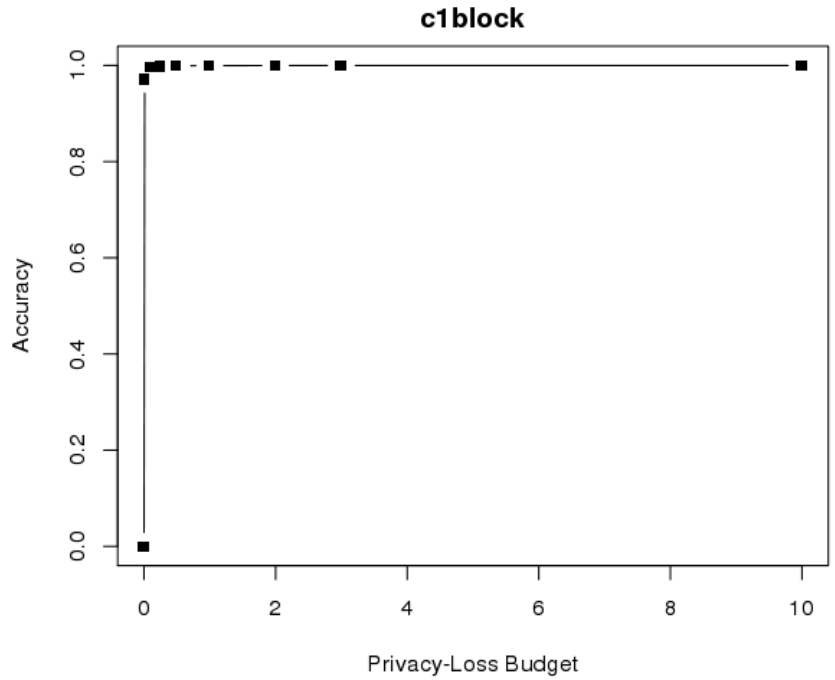
Accuracy v. Privacy Loss for Rhode Island (2010 Census) using the 2018 E2E Test Disclosure Avoidance System

PL94-171 redistricting data at the state level









United States Census Bureau