Disclosure Avoidance for Block Level Data and Protection of Confidentiality in Public Tabulations

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
Census Scientific Advisory Committee
December 6, 2018

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

Slides updated December 4, 2018

Acknowledgments

- The Census Bureau's 2020 Disclosure Avoidance System incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Scientist for Confidentiality and Data Access), Rob Sienkiewicz (ACC Disclosure Avoidance, Center for Enterprise Dissemination), Tamara Adams, Robert Ashmead, Michael Bentley, Stephen Clark, Craig Corl, Aref Dajani, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Edward Porter, Sarah Powazek, Anne Ross, Ian Schmutte, William Sexton, Lars Vilhuber, Cecil Washington, and Pavel Zhuralev.

A Brief History of Differential Privacy at the U.S. Census Bureau

This work was done at Cornell University while Abowd and Vilhuber were on IPA assignments to the Census Bureau. Gehrke is now Technical Fellow at Microsoft. Kifer is now the scientific lead on the 2020 DAS. Machanavajjhala is now a contractual collaborator on the 2020 DAS. Vilhuber is now on IPA assignment to the Census Bureau.

2008

Privacy: Theory meets Practice on the Map

Ashwin Machanavajjhala [†1], Daniel Kifer [†2], John Abowd [#3], Johannes Gehrke [†4], Lars Vilhuber [#5]

[†]*Department of Computer Science, Cornell University, U.S.A.*
[#]*Department of Labor Economics, Cornell University, U.S.A.*

[1]mvnak@cs.cornell.edu [2]dkifer@cs.cornell.edu [3]john.abowd@cornell.edu
[4]johannes@cs.cornell.edu [5]lars.vilhuber@cornell.edu

***Abstract*— In this paper, we propose the first formal privacy analysis of a data anonymization process known as the synthetic data generation, a technique becoming popular in the statistics community. The target application for this work is a mapping program that shows the commuting patterns of the population of the United States. The source data for this application were collected by the U.S. Census Bureau, but due to privacy constraints, they cannot be used directly by the mapping program. Instead, we generate synthetic data that statistically mimic the original data while providing privacy guarantees. We use these synthetic data as a surrogate for the original data.**

We find that while some existing definitions of privacy are inapplicable to our target application, others are too conservative and render the synthetic data useless since they guard against privacy breaches that are very unlikely. Moreover, the data in our target application is sparse, and none of the existing solutions are tailored to anonymize sparse data. In this paper, we propose solutions to address the above issues.

I. Introduction

In this paper, we study a real-world application of a privacy preserving technology known as synthetic data generation. We present the first formal privacy guarantees (to the best of our knowledge) for this application. This paper chronicles the challenges we faced in this endeavour. The target application is based on data developed by the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD). By combining various Census datasets it is possible to construct a table Commute_Patterns with schema *(id,origin_block, destination_block)* where each row represents a worker. The attribute *id* is a random number serving as a key for the table, *origin_block* is the census block in which the worker lives, and *destination_block* is where the worker works. An origin block o corresponds to a destination block d if there is a tuple with *origin_block* o and *destination_block* d. The goal is to plot points on a map that represent commuting patterns for the U.S. to such a mapping application. An anonymized version must be used instead.

The algorithm used to anonymize the data for the above mapping application is known as the synthetic data generation [1], which is becoming popular in the statistical disclosure limitationg community. The main idea behind synthetic data generation is to build a statistical model from the data and then to sample points from the model. These sampled points form the synthetic data, which is then released instead of the original data. While much research has focused on deriving the variance and confidence intervals for various estimators from synthetic data [2], [3], there has been little research on deriving formal guarantees of privacy for such an approach (an exception is [4]).

Much recent research has focused on deriving formal criteria for privacy. These include general notions of statistical closeness [5], variants of the notions of *k-anonymity* [6] and *ℓ-diversity* [7], [8], [9], [10], [11], (ρ_1, ρ_2)*-privacy* [12], and variants of *differential privacy* [13], [14]. However, we found that apart from the differential privacy criterion [13], none of the other privacy conditions applied to our scenario.

Picking an off-the-shelf synthetic data generation algorithm and tuning it to satisfy the differential privacy criterion was unsatisfactory for the following reasons. First, in order to satisfy the differential privacy criterion, the generated synthetic data contained little or no information about the original data. We show that this is because differential privacy guards against breaches of privacy that are very unlikely.

Next, no deterministic algorithm can satisfy differential privacy. Randomized algorithms can (albeit with a very small probability) return anonymized datasets that are totally unrepresentative of the input. This is a problem, especially, when we want to publish a single or only a few, anonymous versions of

https://onthemap.ces.census.gov

OnTheMap

LEHD Home Help and Documentation Reload Text-Only

Start | Base Map | Selection

Save | Load | Feedback | Previous Extent | Hide Tabs

Welcome to OnTheMap!

Start an analysis by using one of the tools below (Search, Import Geography, or Load .OTM file). Hover over the Help icons located throughout the application to see Help tips for using specific functionality. Sections in the control panel can be collapsed or opened by clicking the section title.

2015 Data Now Available (09/25/2017)

Search

Search

Search All Names

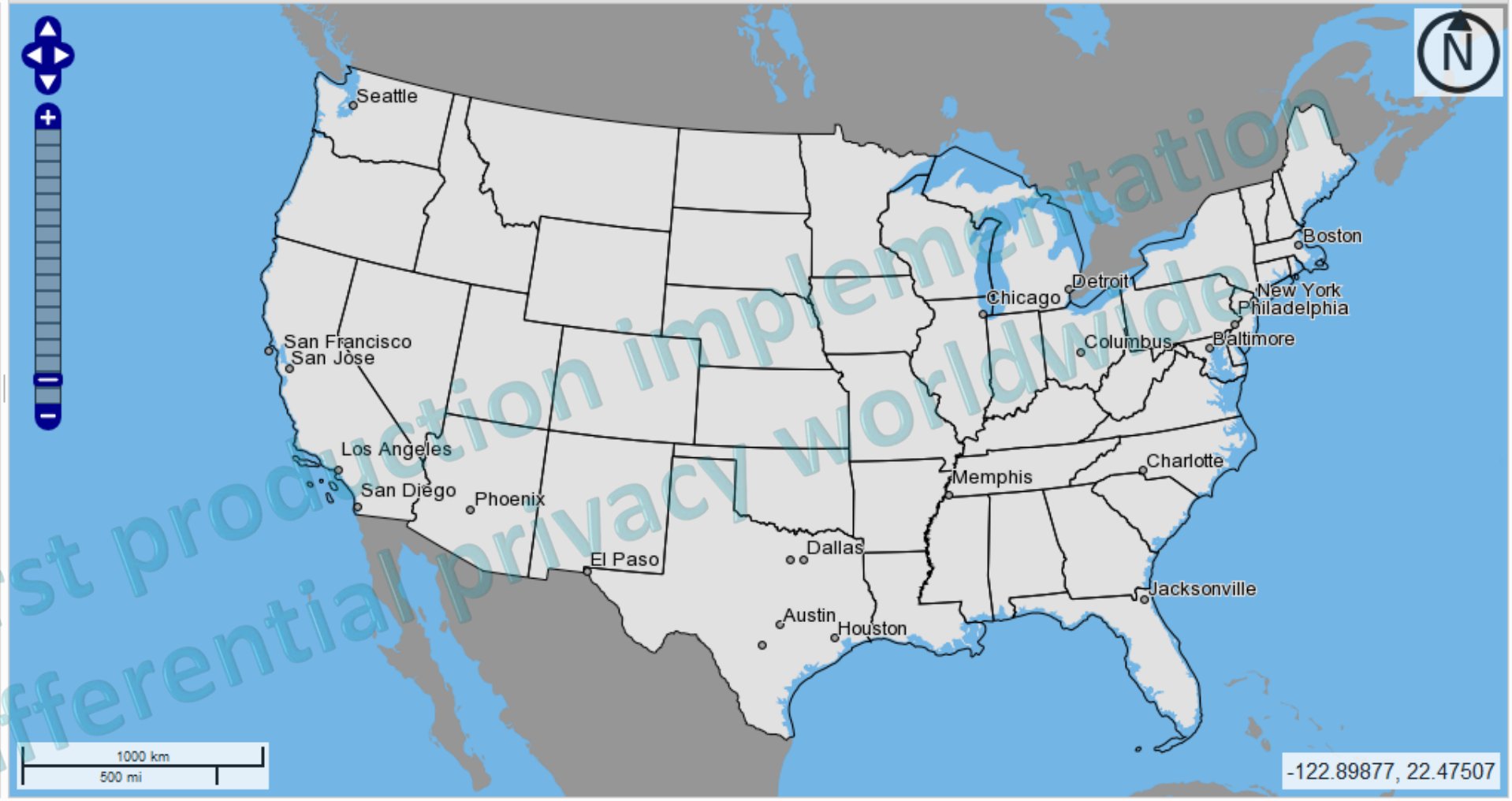
Import Geography

Import from KML
Import from SHP
Import from GPS

Load .OTM File

Click the "Load" button below to load a .OTM file.

Load



2008: First production implementation of differential privacy worldwide

Privacy Policy | 2010 Census | Data Tools | Information Quality | Product Catalog | Contact Us | Home

Source: U.S.Census Bureau, Center for Economic Studies | e-mail: CES.OnTheMap.Feedback@census.gov

Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

https://onthemap.ces.census.gov

OnTheMap

LEHD Home | Help and Documentation | Reload | Text-Only

Start | Base Map | Selection | Results

Distance/Direction Analysis
Work to Home

Display Settings

Labor Market Segment Filter: All Workers

Year: 2015

Map Controls

Color Key

Thermal Overlay ✓

Point Overlay ✓

Selection Outline ✓

Identify | Zoom to Selection

Clear Overlays | Animate Overlays

Report/Map Outputs

Detailed Report

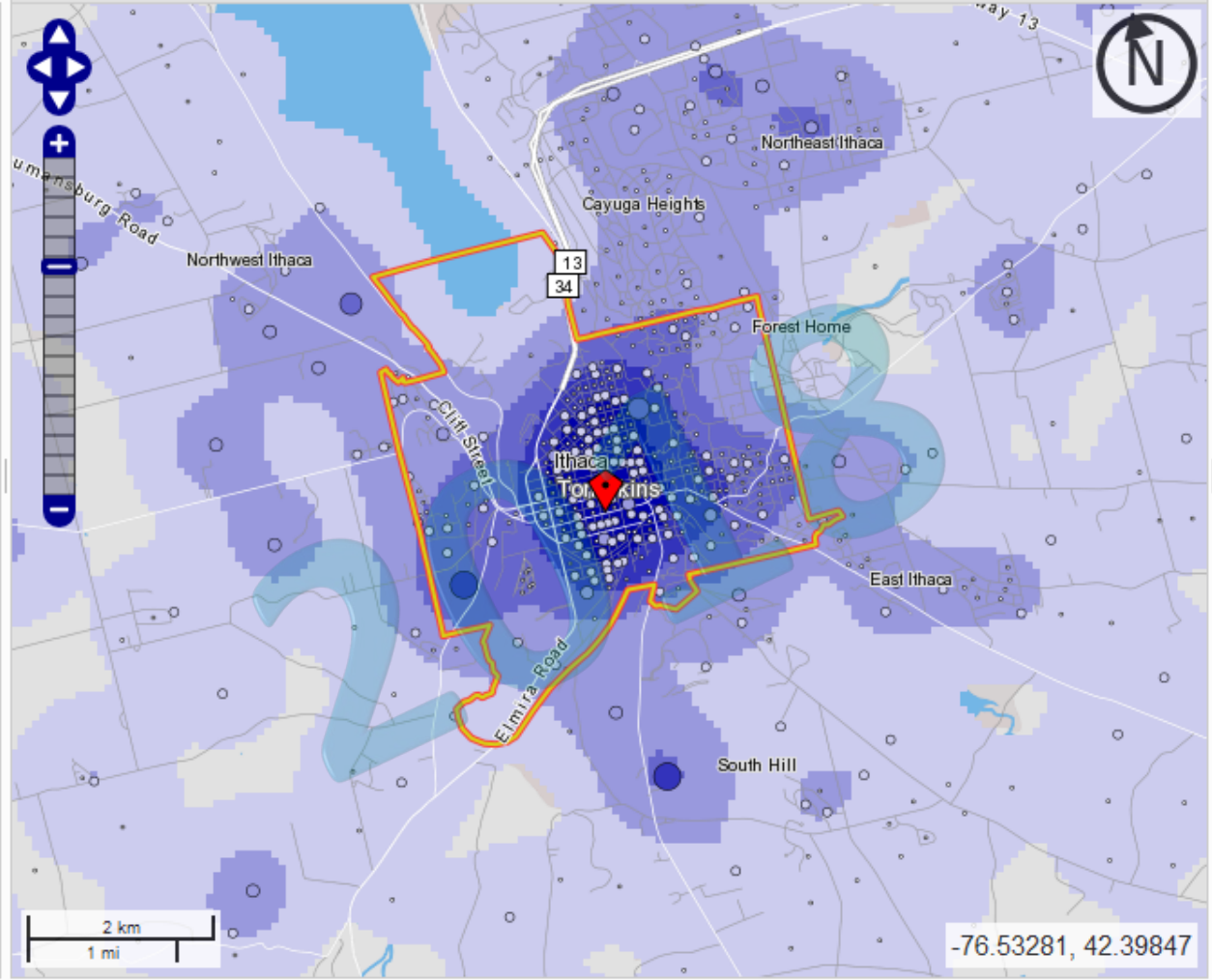
Export Geography

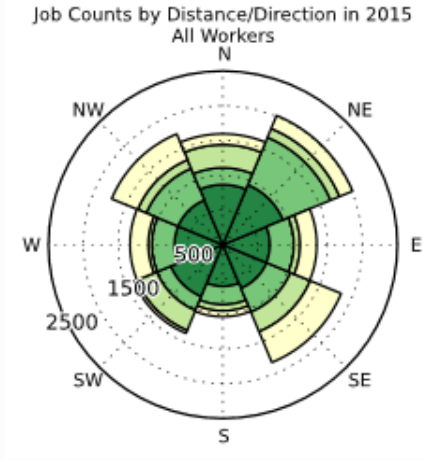
Print Chart/Map

Legends

Change Settings

Save | Load | Feedback | Previous Extent | Hide Tabs | Hide Chart/Report





View as Radar Chart

Jobs by Distance - Work Census Block to Home Census Block

	2015 Count	2015 Share
Total Primary Jobs	12,260	100.0%
Less than 10 miles	5,949	48.5%
10 to 24 miles	2,987	24.4%
25 to 50 miles	1,451	11.8%
Greater than 50 miles	1,873	15.3%

Privacy Policy | 2010 Census | Data Tools | Information Quality | Product Catalog | Contact Us | Home

Source: U.S.Census Bureau, Center for Economic Studies | e-mail: CES.OnTheMap.Feedback@census.gov

Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Census Data Stewardship

Our Overall Approach to Maintain Public Trust

This slide is from the August 3, 2018 Program Management Review (PMR) for the 2020 Census. 2020 PMRs are quarterly public presentations of the state of readiness of the decennial census.

Phase 1 — Create and Reinforce Culture

- **Create & Reinforce** — Create Stewardship Culture and Enforce Laws and Policies

Phase 2 — Implement Secure Systems

- **Collaborate** — With Partners in Federal Government and Industry
- **Coordinate** — With Federal Government and Industry using NIST Cybersecurity Framework
- **Design** — That Contain Cyber Threats and Sustain Response Services to Maintain Public Trust
- **Monitor** — Threats to Protect Against and Detect Cyber Issues

Phase 3 — Secure Data Collection and Dissemination

- **Collect** — Data Securely with Encryption Everywhere
- **Isolate** — Data as Soon as it is Submitted
- **Process & Validate** — Data with People, Processes, and systems
- **Protect Confidentiality** — Differential Privacy

Phase 4 — Monitor and Respond

- **Test & Monitor** — Processes and Technology for cyber issues
- **Communicate** — With Public through Communications Campaign

2018: First large-scale production implementation of central differential privacy worldwide

Database Reconstruction

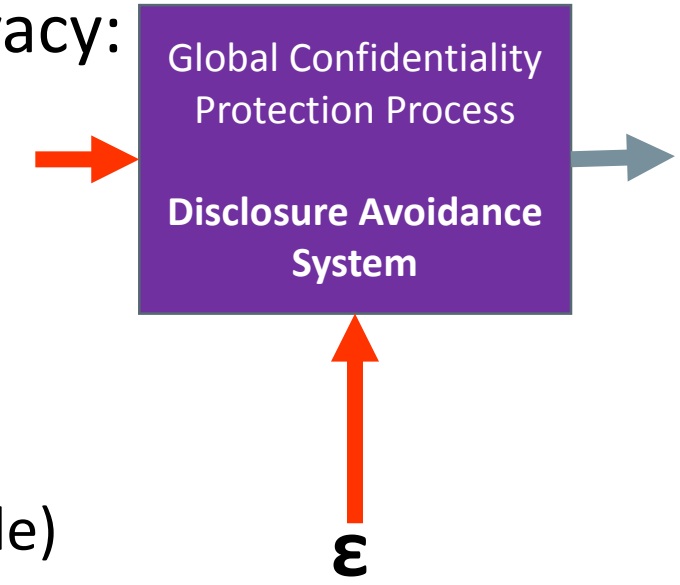
Internal Experiments Using the 2010 Census

- Confirm that the micro-data from the confidential 2010 Hundred-percent Detail File (HDF) can be accurately reconstructed from PL94 + balance of SF1
- While there is a reconstruction vulnerability, the risk of re-identification using only person-level is apparently still relatively small
- But, experiments were not done at the household-level, so there is no re-identification risk quantification possible there
- Experiments have led to the declaration that reconstruction of Title 13-sensitive data is an issue, no longer a risk
- Strong motivation for the adoption of differential privacy for the 2018 End-to-End Census Test and 2020 Census

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*



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

Hierarchical and Table Consistency

- Tables are consistent at all levels of geography and for all coarsening of detail
- The DP noise is infused into measurements (contingency tables) at each level of geography (uses closure under composition)
  - Global privacy-loss budget is shared across all measurements with correct accounting
  - Geometric mechanism noise (discrete version of Laplace noise)
- Resulting tables are post-processed (uses robustness to post-processing)
  - To satisfy all invariants
  - To contain only non-negative value
  - To add up properly along all dimensions
  - The large-scale optimization problem is solved partially analytically and partially using Gurobi
- One set of consistent micro-data is produced that tabulates exactly to the post-processed tables
- These micro-data are sent to the production tabulation system

Invariants

- An invariant is an output of the disclosure avoidance system that exactly matches the query answer in the confidential data
  - Shorthand: no disclosure avoidance was applied
- Public documents confirm that total population, voting-age population, number of housing units, number of occupied housing units, and number and type of group quarters were all invariants at the block level in 2000 and 2010 Census publications. (These, plus race and ethnicity, were also invariant in the 1990 Census.)
- We now understand that invariants seriously compromise the confidentiality protect (differential privacy or traditional disclosure avoidance)

How Invariants Compromise Confidentiality

- For traditional methods: database reconstruction is much easier
- For differential privacy:
  - Secrets like (Alice lives in block A and Alice is a citizen) do not get the differential privacy protection if *either* block population or citizen population is an invariant
  - Unrelated secrets (those not involving invariants at all) can also be compromised
  - When the differential privacy protection does not apply, the meaningful bounds on the accuracy of inferences about Alice's residence or citizenship are much more complex than when there are no invariants
- Eliminating invariants eliminates this problem, allows for more efficient use of the privacy-loss budget

Hierarchical Consistency + Invariants

- The combination makes the post-processing algorithms extremely complicated
- There are integer, equality and inequality constraints in the equation system
- Proofs of existence are not available for every invariant and/or consistency configuration
- When existence fails empirically, the system reverts to an approximate solution for a configuration that has the smallest feasible subset of the problem

Asymptotic Consistency

- The system should have the property that when the global privacy-loss budget goes to infinity, the solution goes to the exact answer to the query workload (perfect accuracy)
- That is not automatic for the top-down algorithm
- To achieve asymptotic consistency, the most general histogram must be estimated at every level of geography with a proportional allocation of the privacy-loss budget
- As ε grows, the accuracy of the most general histogram also improves at every level of the geographic hierarchy

Existence Proofs

- Differential privacy can always be applied to every measurement needed by the top-down algorithm
- However, the equation system used for the post-processing need not always have a solution consisting entirely of non-negative integers
- We have existence proofs for some configurations
- In general, when there are inequality constraints, existence is not guaranteed
- Approximate solutions are computed over a minimal subset
- Rate of approximation will be in the error metrics

Workload Optimization

- Most of the work for the 2018 End-to-End Census Test was done at the person-level, and for the PL94-171 (redistricting) workload
- There are many more person-level tables in the proposed Summary File 1
- There are also household-level tables (based on householder/person 1), and joins of household and person tables (counts of persons by characteristics of the householder)
  - The household tables will require workload optimization (probably using the high-dimensional matrix method https://arxiv.org/abs/1808.03537)
  - The joins will require other algorithmic extensions (probably using hierarchical counts-of-counts technique https://arxiv.org/abs/1804.00370)

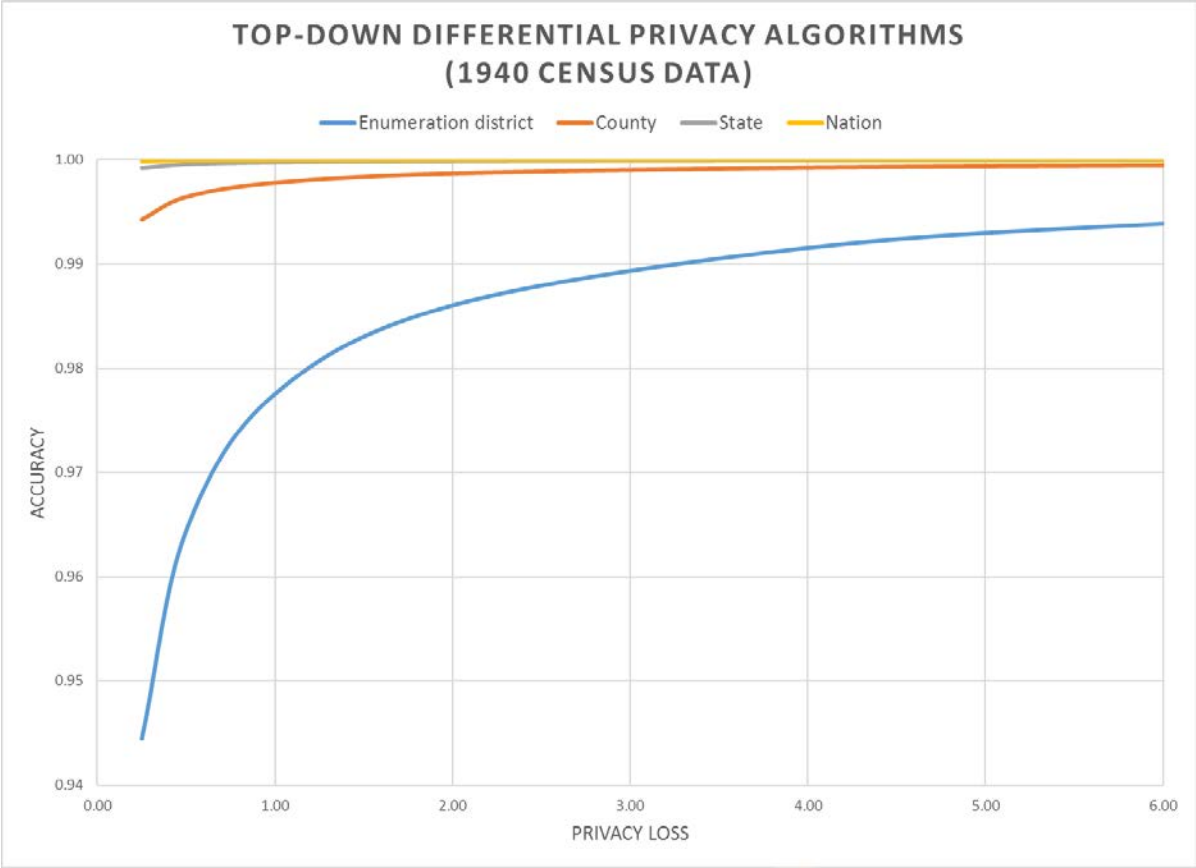
Summary of Important Technical Details

- Central differential privacy implementation with a controlled total privacy-loss budget
- Relevant algorithmic definition is bounded ε-differential privacy (total population of the largest geographic area is public)
- Semantic privacy guarantee is $[-2\varepsilon, 2\varepsilon]$ by properties of bounded differential privacy
  - Complicated by occupied household invariant that was not removed
- All algorithms, code, and parameter values will be released with the test files for the 2018 End-to-End Census Test
- All parameter and invariant settings to be reviewed based on feedback from the E2E test files and released algorithms

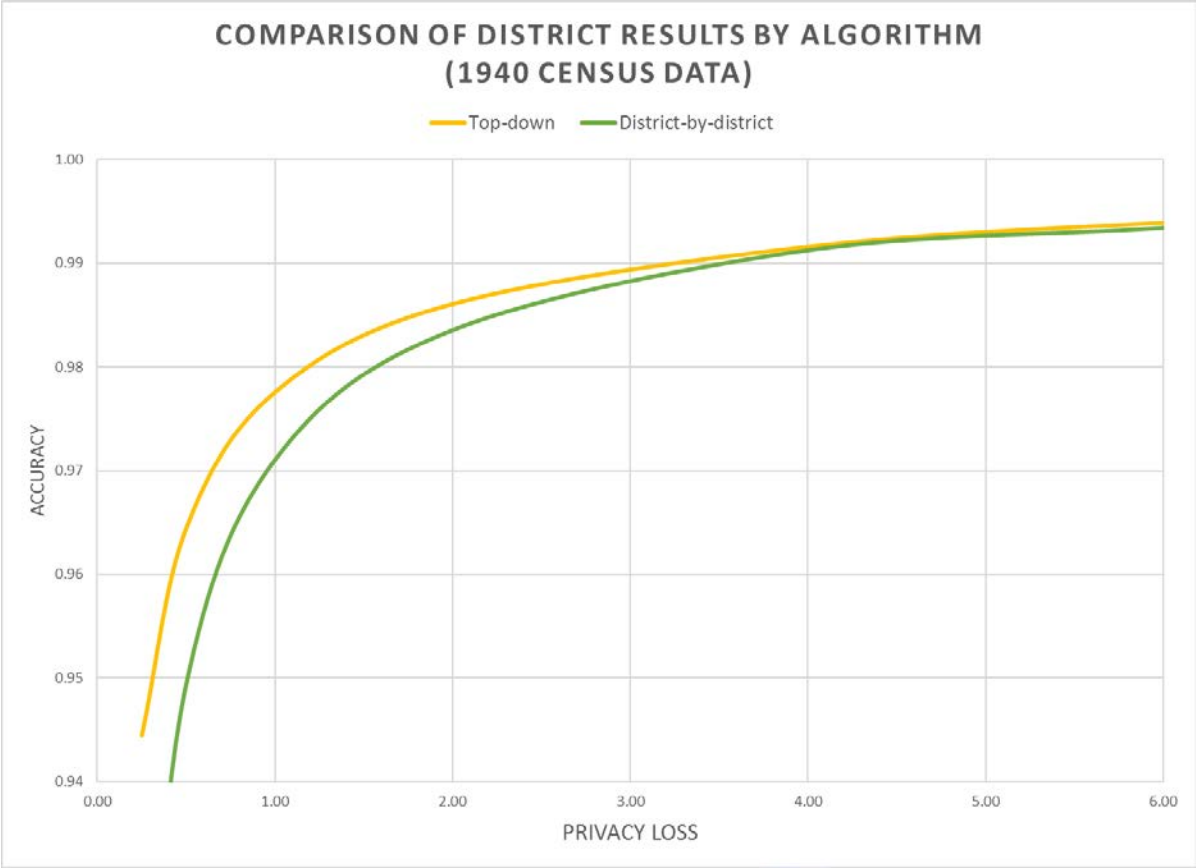
Algorithm Comparison Using Public 1940 Census Data

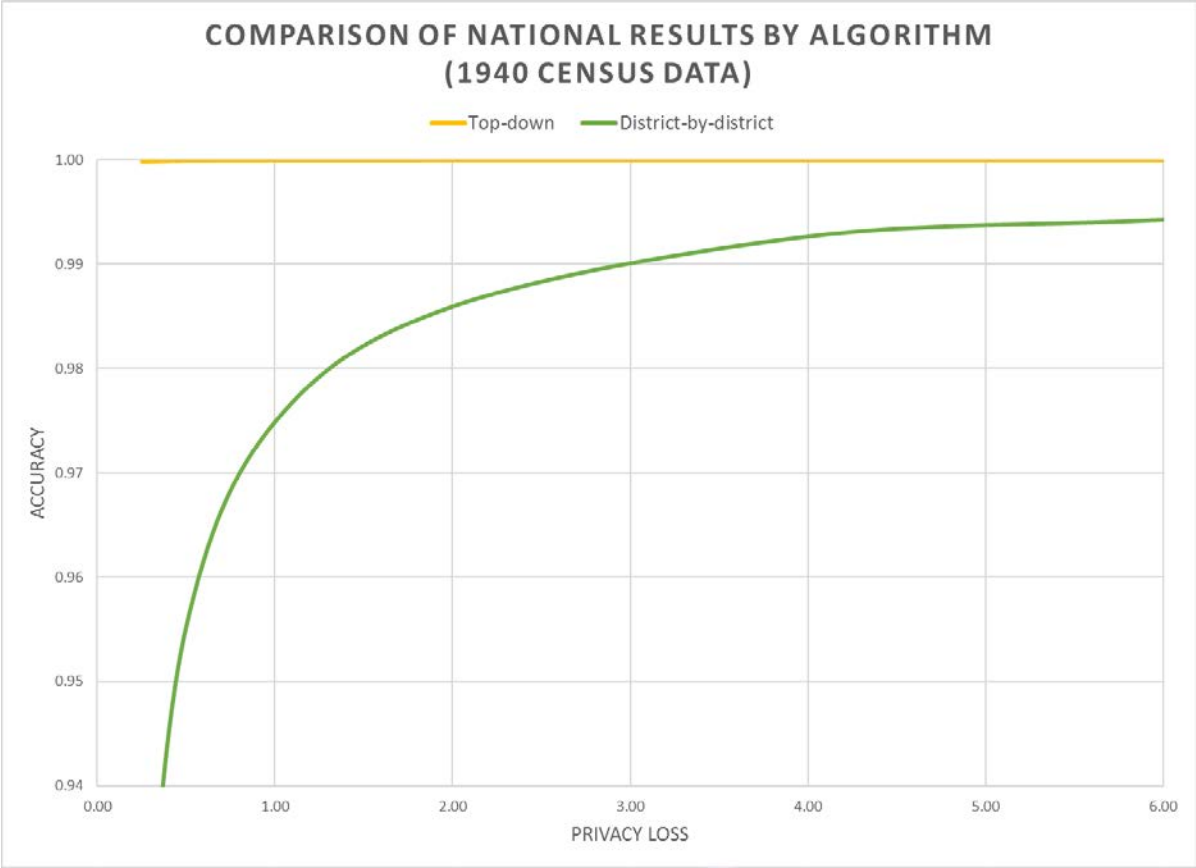
Two Candidate Algorithms

- 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
ACCURACY
1.00
0.99
0.98
0.97
0.96
0.95
0.94
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
ACCURACY
1.00
0.99
0.98
0.97
0.96
0.95
0.94
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
ACCURACY
1.00
0.99
0.98
0.97
0.96
0.95
0.94
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
ACCURACY
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS

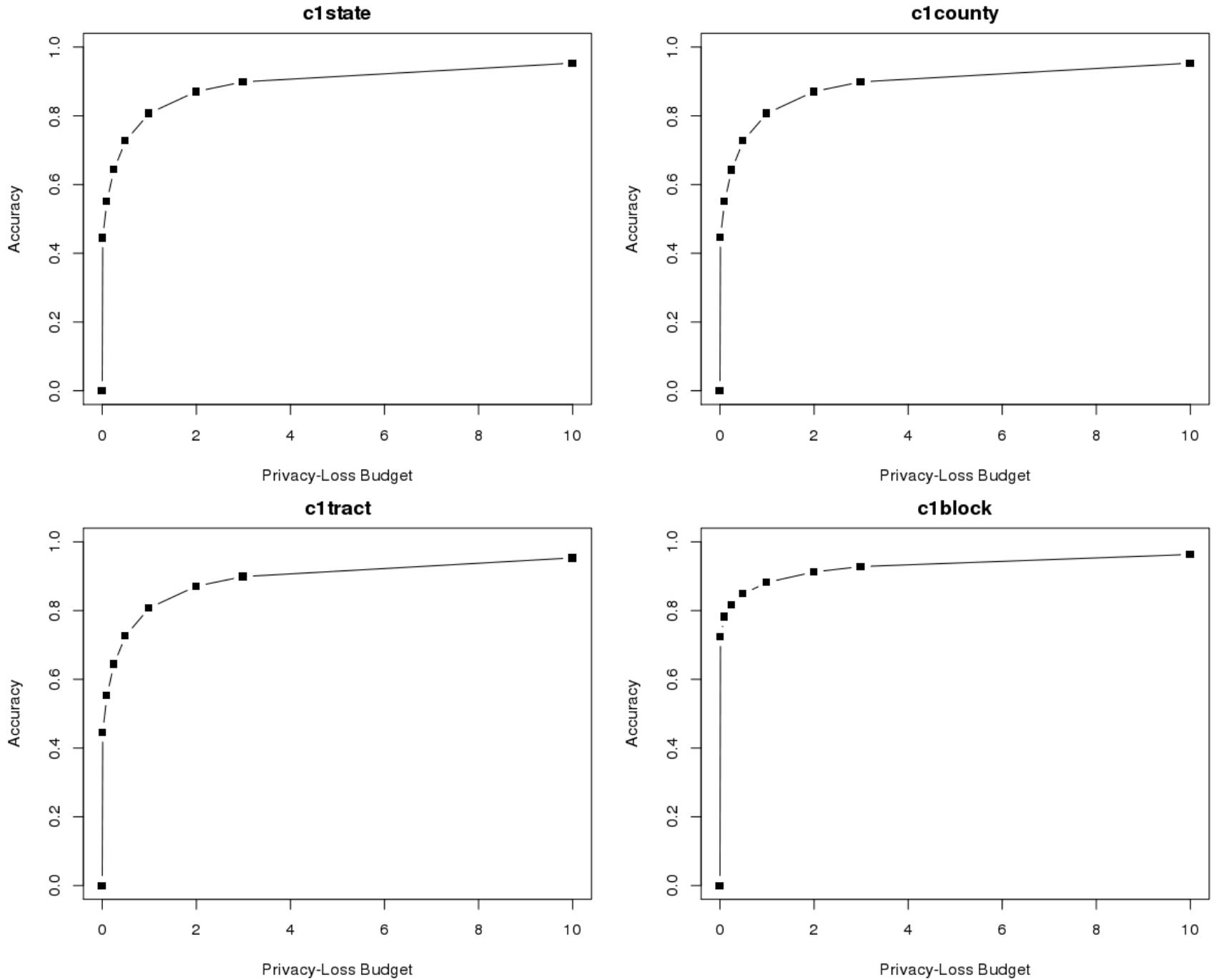
Analyses Supporting the 2018 End-to-End Census Test

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

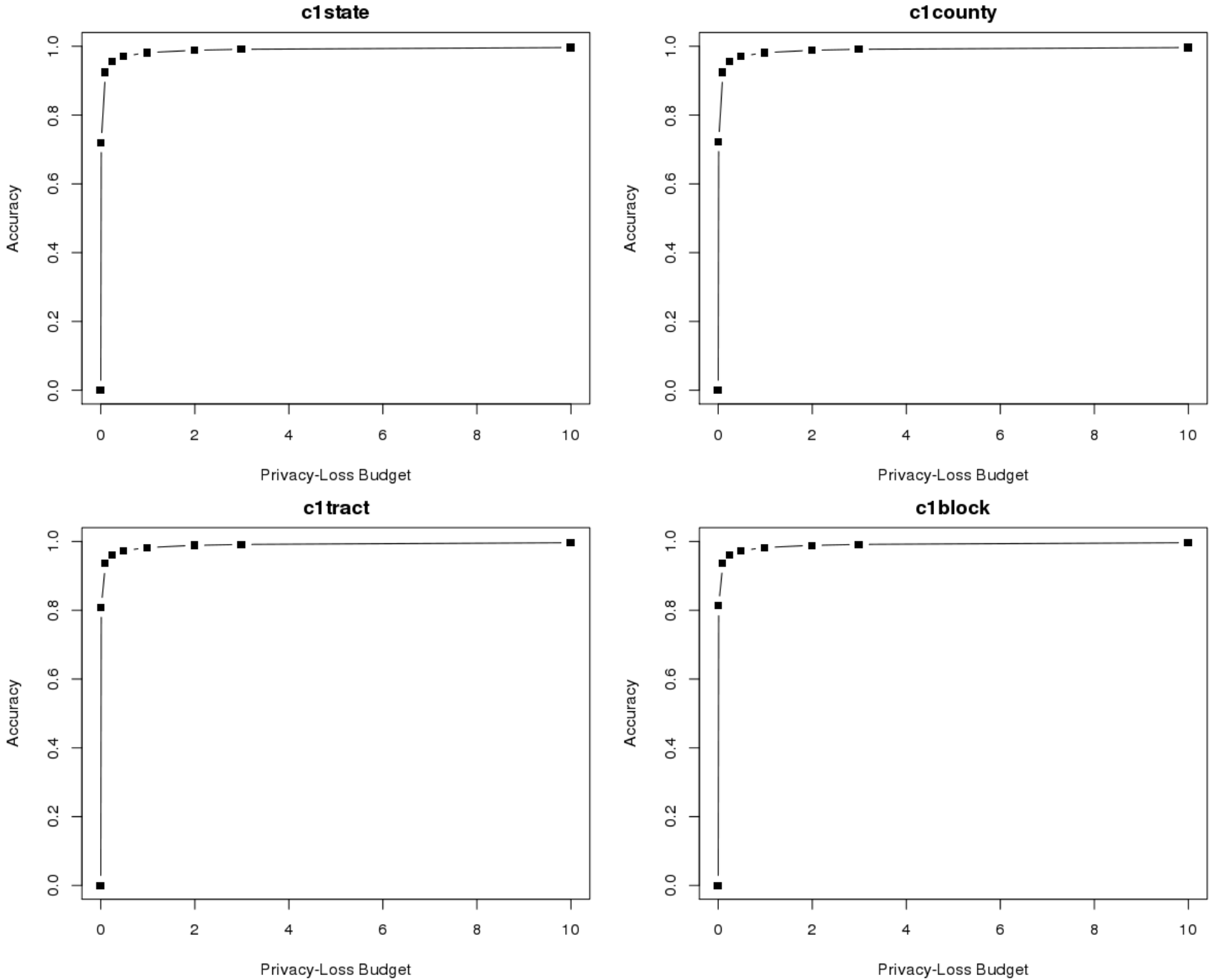
PL94-171 redistricting data at the block level



United States Census Bureau

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

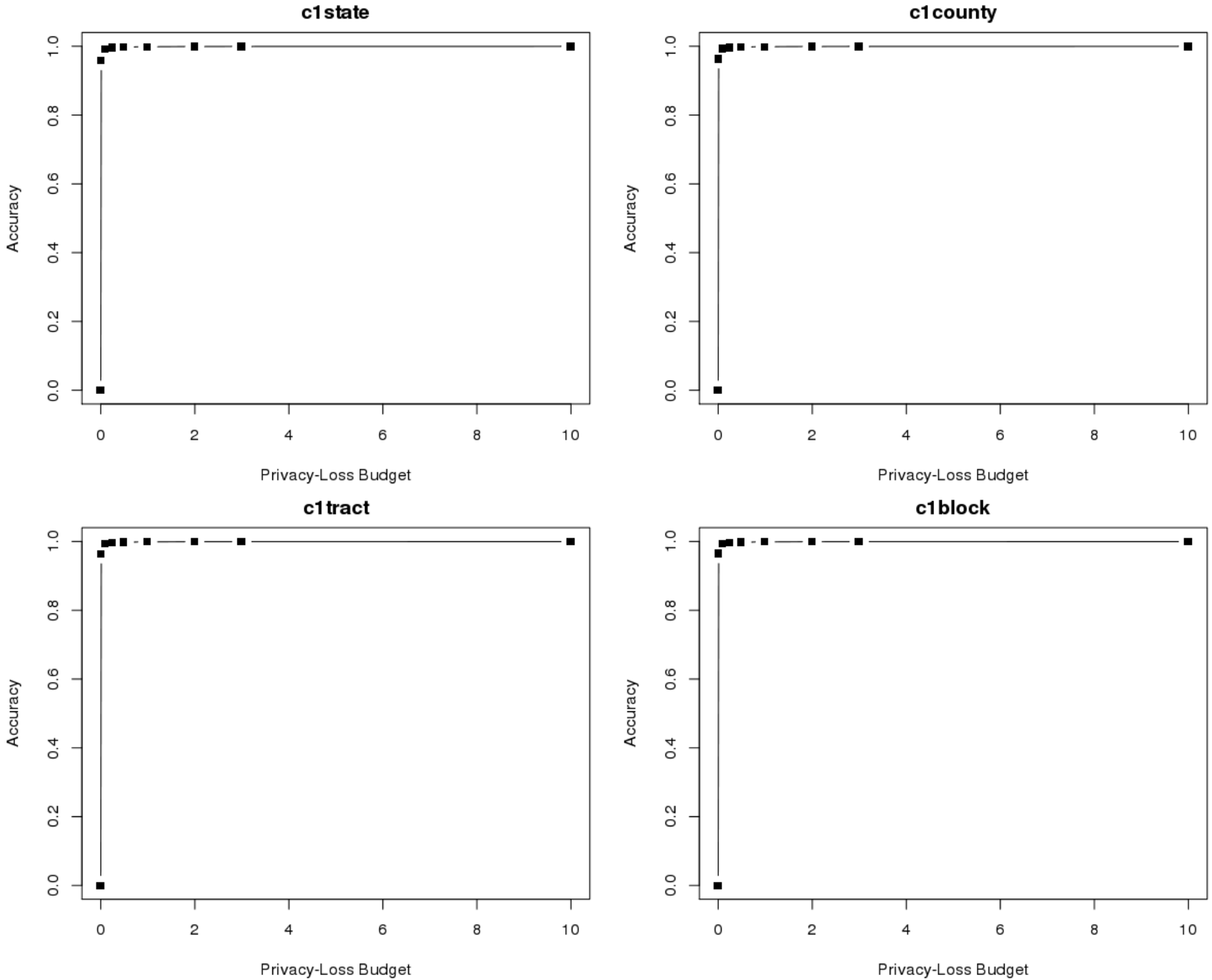
PL94-171 redistricting data at the tract level



United States Census Bureau

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

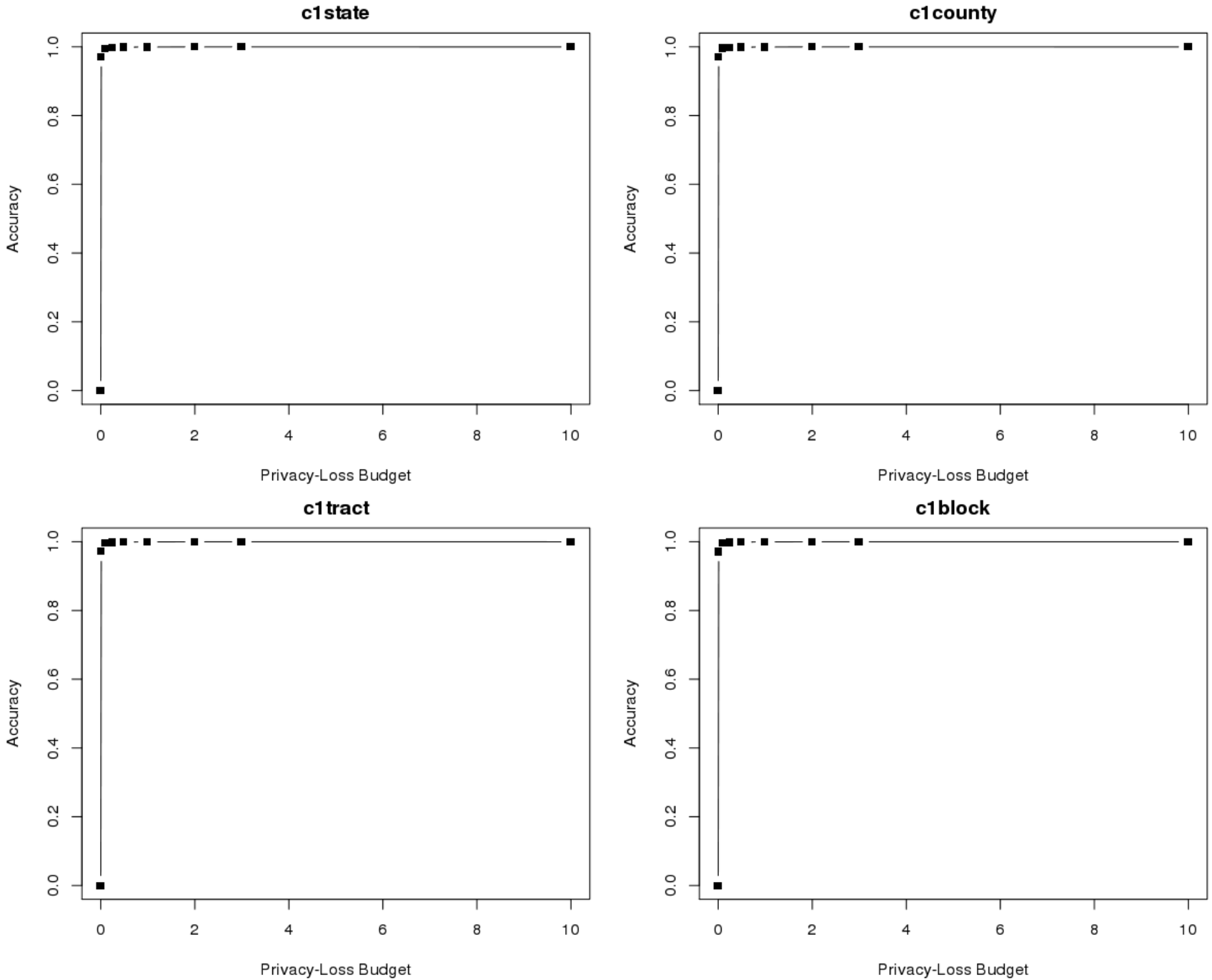
PL94-171 redistricting data at the county level



United States Census Bureau

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

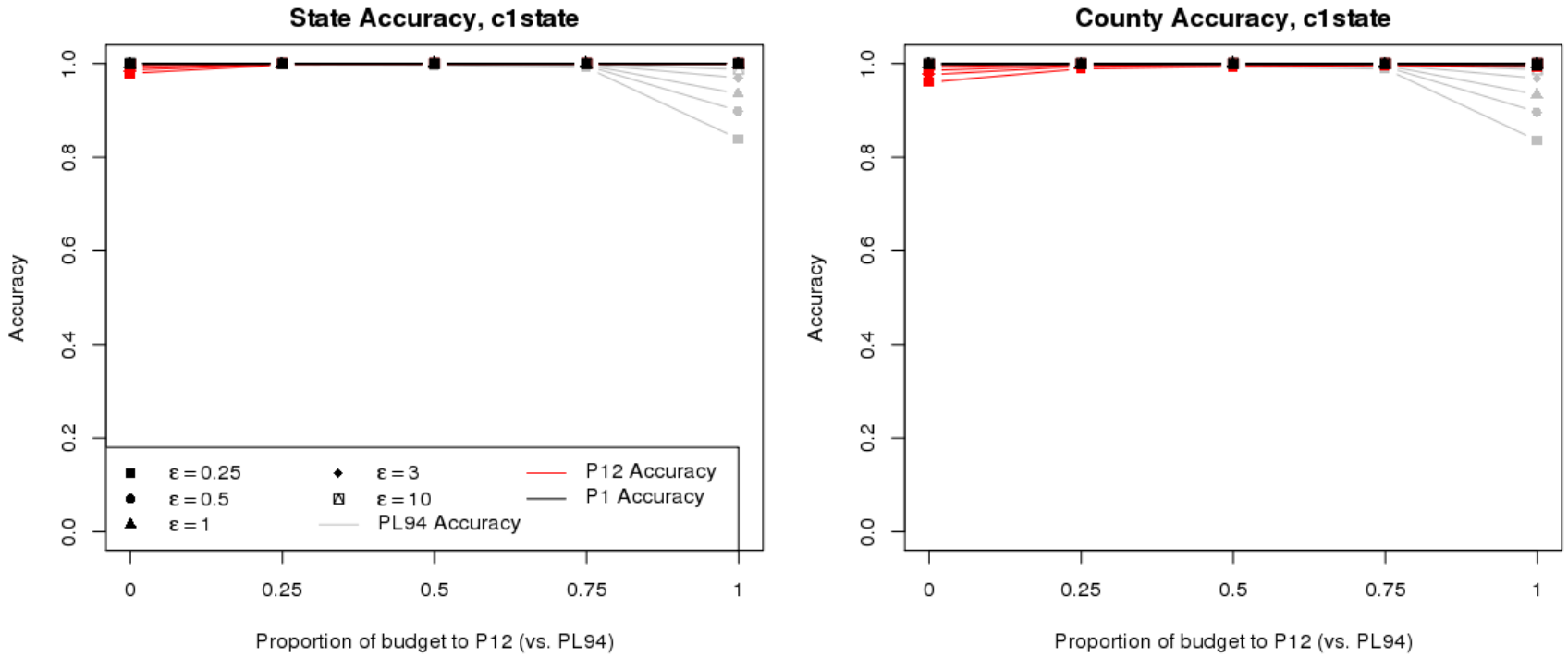
PL94-171 redistricting data at the state level

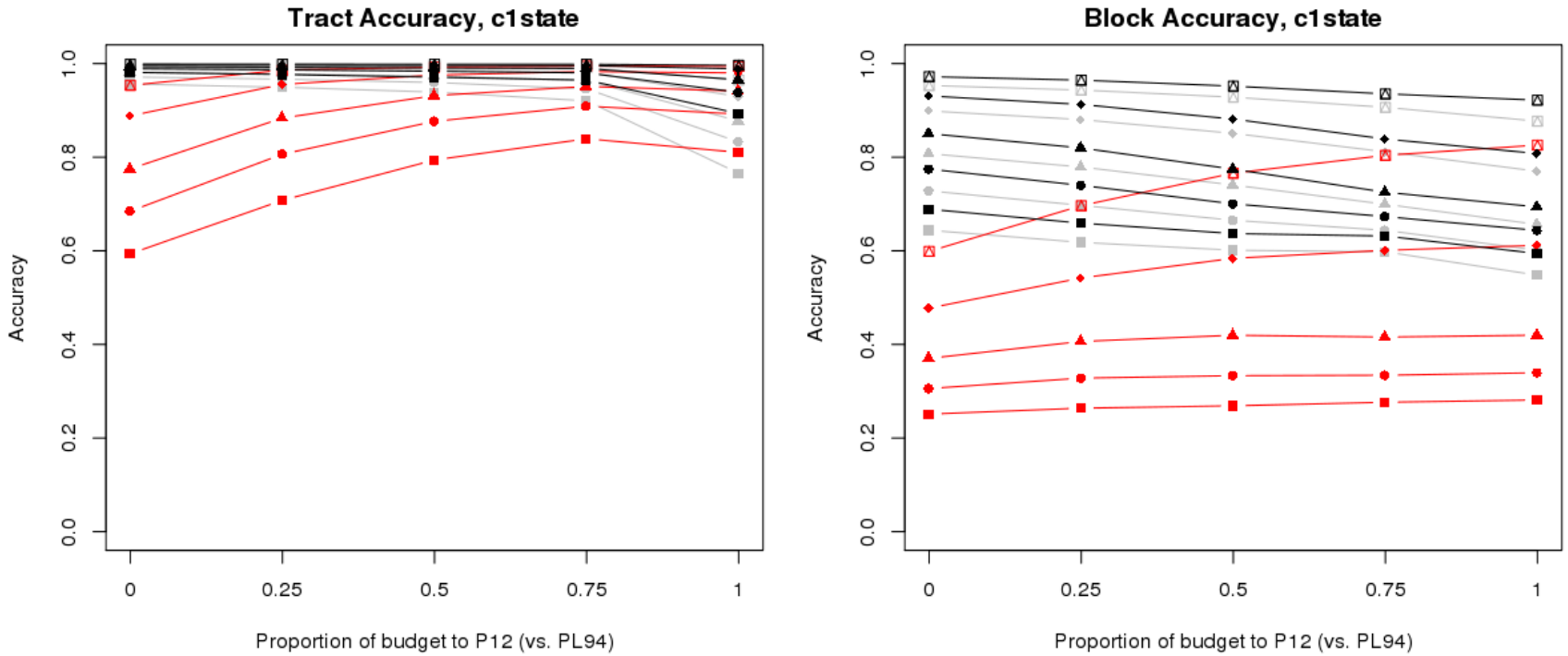


United States Census Bureau

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

PL94-171: redistricting data, SF1: P12 age x sex data and P1 population data



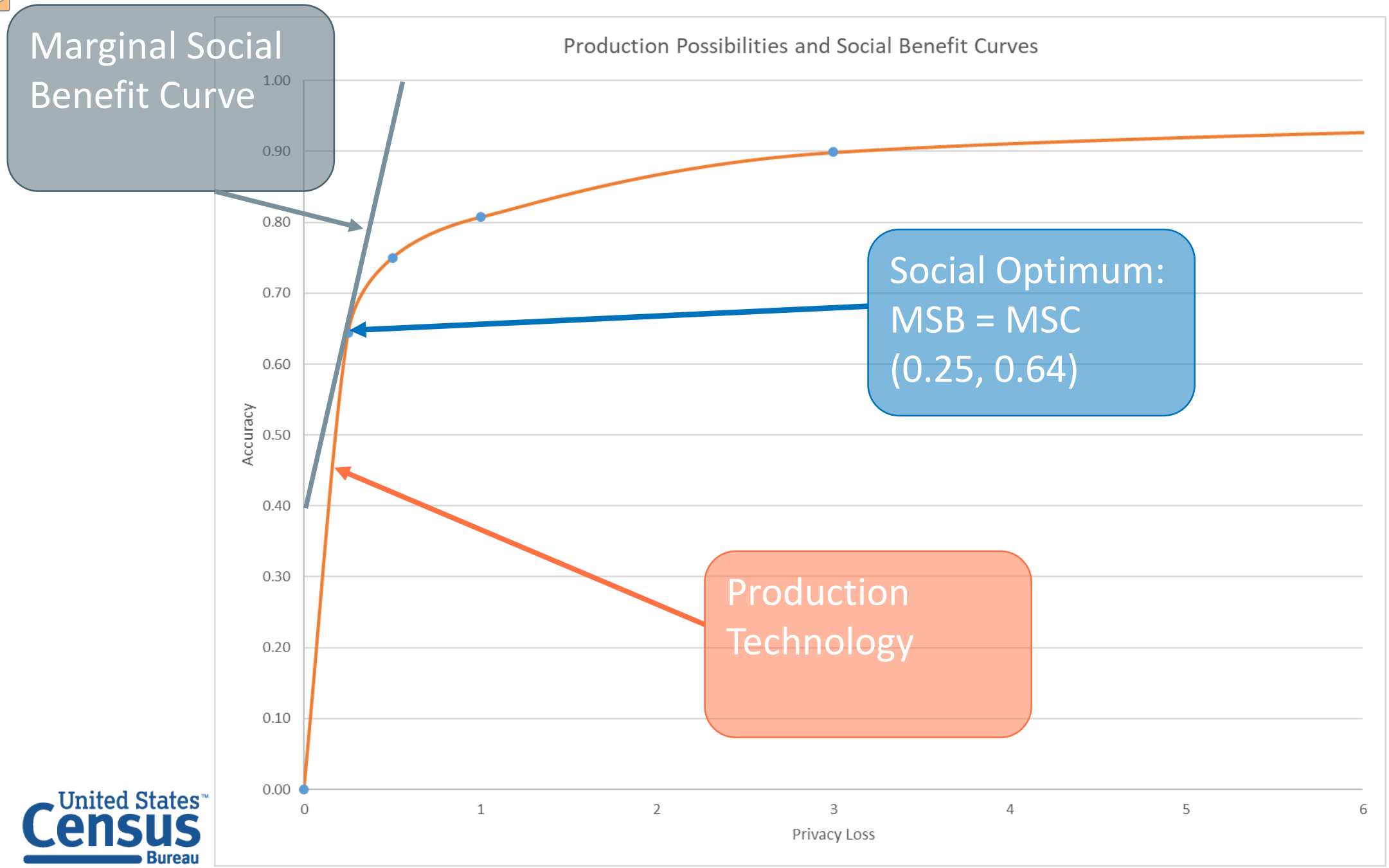


United States Census Bureau

Managing the Tradeoff

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



United States Census Bureau

But the Choice Problem for Redistricting Tabulations Is More Challenging

- 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 detailed race, ethnicity and citizenship

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

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. 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.