Staring-Down the Database Reconstruction Theorem

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
Joint Statistical Meetings, Vancouver, BC, Canada
July 30, 2018

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

Acknowledgments and Disclaimer

- The opinions expressed in this talk are the my own and not necessarily those of the U.S. Census Bureau
- The application to the Census Bureau's 2020 publication system incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Scientist for Confidentiality and Data Access), Tamara Adams, Robert Ashmead, Michael Bentley, Stephen Clark, Aref Dajani, Jason Devine, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Gerome Miklau, Brett Moran, Edward Porter, Anne Ross, and Lars Vilhuber [link to the September 2018 Census Scientific Advisory Committee presentation]
- Parts of this talk were supported by the National Science Foundation, the Sloan Foundation, and the Census Bureau (before and after my appointment started)

Outline

- Database reconstruction is an issue, not a risk
- Examples from the 2010 Census of Population and Housing
- The risks in conventional statistical disclosure limitation
- 2018 End-to-End Test (block-by-block)
- 2020 Census (top down)
- How to think about the social choice problem of setting ε

Database Reconstruction

2003:
Database Reconstruction

ABSTRACT

We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an n-bit string $d_1, .., d_n$, with a query being a subset $q \subseteq [n]$ to be answered by $\sum_{i \in q} d_i$. Our main result is a polynomial reconstruction algorithm of data from noisy (perturbed) subset sums. Applying this reconstruction algorithm to statistical databases we show that in order to achieve privacy one has to add perturbation of magnitude $\Omega(\sqrt{n})$. That is, smaller perturbation always results in a strong violation of privacy. We show that this result is tight by exemplifying access algorithms for statistical databases that preserve privacy while adding perturbation of magnitude $\tilde{O}(\sqrt{n})$.

For time-$\mathcal{T}$ bounded adversaries we demonstrate a privacy-preserving access algorithm whose perturbation magnitude is $\approx \sqrt{\mathcal{T}}$.

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Revealing Information while Preserving Privacy

Irit Dinur Kobbi Nissim*
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
{iritd,kobbi}@research.nj.nec.com

ABSTRACT

We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an n-bit string $d_1, .., d_n$, with a query being a subset $q \subseteq [n]$ to

the hospital is obliged to keep the privacy of its patients, i.e. leak no medical information that could be related to a specific patient. The hospital needs an access mechanism to the database that allows certain 'statistical' queries to be answered, as long as they do not violate the privacy of any single patient.

*Work partly done when the author was at DIMACS, Rutgers University, and while visiting Microsoft Research Silicon Valley Lab.

PODS 2003, June 9-12, 2003, San Diego, CA.

One simple tempting solution is to remove from the database all 'identifying' attributes such as the patients' names and social security numbers. However, this solution is not enough to protect patient privacy since there usually exist other means of identifying patients, via *indirectly identifying* attributes may

Viewing

approaches taken into three main categories: (i) query restriction, (ii) data perturbation, and (iii) output perturbation. We give a brief review of these approaches below, and refer the reader to [2] for a detailed survey of the methods and their weaknesses.

Query Restriction. In the query restriction approach, queries are required to obey a special structure, supposedly to prevent the querying adversary from gaining too much information about specific database entries. The limit of this approach is that it allows for a relatively small number of queries.

A related idea is of query auditing [7], i.e. a log of the queries is kept, and every new query is checked for possible compromise, allowing/disallowing the query accordingly.

[1]A patient's gender, approximate age, approximate weight, ethnicity, and marital status – may already suffice for a complete identification of most patients in a database of a thousand patients. The situation is much worse if a relatively 'rare' attribute of some patient is known. For example, a patient having Cystic Fibrosis (frequency $\approx 1/3000$) may be uniquely identified within about a million patients.

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 $\sqrt{N}$

2010 Census of Population: Summary

Total population	308,745,538
Household population	300,758,215
Group quarters population	7,987,323
Households	116,716,292

2010 Census: High-level Database Schema

Variables	Distinct values
Habitable blocks	10,620,683
Habitable tracts	73,768
Sex	2
Age	115
Race/Ethnicity (OMB Categories)	126
Race/Ethnicity (SF2 Categories)	600
Relationship to person 1	17
National histogram cells (OMB Ethnicity)	492,660

2010 Census: Published Statistics

Publication	Released counts (including zeros)
PL94-171 Redistricting	2,771,998,263
Balance of Summary File 1	2,806,899,669
Summary File 2	2,093,683,376
Public-use micro sample	30,874,554
Lower bound on published statistics	7,703,455,862
Statistics/person	25

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

Internal Experiments Using the 2010 Census

- Confirm that the confidential micro-data from the hundred percent detail file can be reconstructed quite accurately from PL94 + balance of SF1
- While we've determined there is a vulnerability, the risk of re-identification is small
- Experiments are at the person level, not household
- 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

Examples from the 2010 Census: PL94

- From PL94-171 (redistricting data) block level:
  - P1 Race
    - Universe: total population
    - OMB race categories ($2^6 - 1 = 63$)
  - P2 Hispanic or Latino, and not Hispanic by Race
    - Universe: total population
    - Hispanic ethnicity (2) x OMB race categories (63)
  - P3 Race for the Population 18 Years and over
    - Universe: total population age 18 years and over
    - OMB race categories (63)
  - P4 Hispanic or Latino, and not Hispanic or Latino by Race for the Population 18 Years and Over
    - Universe: total population age 18 years and over
    - Hispanic ethnicity (2) x OMB race categories (63)
  - Note: implies 2 age categories 0-17, 18+

Examples from the 2010 Census: SF1

- From SF1 (summary file 1) block level:
  - P12 Sex by Age
    - Universe: total population
    - Sex (2) by Age in five-year groups (0-4, 5-9, …, 80-84, 85+; 23 groups)
  - P12A-I Sex by Age iterated over OMB race groups (A-G) and Hispanic Origin (H, I)
  - P14 Sex by Age for the Population under 20 years
    - Universe: total population under 20 years old
    - Sex (2) by Age (single-year age 0, 1, 2, …, 19; 20 groups)
- SF1 tract level
  - PCT12 Sex by Age
    - Universe: total population
    - Sex (2) by Age in single years (0, 1, 2, …, 99, 100-104, 105-109, 110+; 103 groups
  - PCT12A-O Sex by Age iterated over OMB race groups (7) x Hispanic Origin (2)

Confidential Record Structure

- Confidential data for the 2010 tabulations
  - Census tract + block geocode (15 digits)
  - Sex (male, female)
  - Age (0, …, 114+; 115 categories)
  - Hispanic or Latino origin (yes/no)
  - White (yes/no)
  - Black or African American (yes/no)
  - Asian (yes/no)
  - American Indian or Alaska Native (yes/no)
  - Native Hawaiian and Other Pacific Islander (yes/no)
  - Some other race (yes/no)
  - Note: race categories White, …, Some other race can be chosen multiply in any combination, but all cannot be no; 63 unique categories

Reconstruction Equation System

- For each of 10,620,683 habitable blocks and 73,768 habitable tracts:
  - Record sample space 2 x 115 x 2 x 63 = 28,980 unique combinations
  - Counts in PL94 tables P1-P4 and SF1 tables P1, P6, P7, P9, P11, P12, P12A-I, P14, PCT12, PCT12A-O provide constraints
  - Margins of tables for total population and voting age population are exact (as per public documentation on PL94-171 and SF1)
  - Only household-level record swapping was used; implies that zeros are unprotected except as swapping relocates them by geography (again, from public documentation on PL94-171 and SF1)

Solving the Equation System I

- Stratify by block within tract:
  - Population counts and voting-age population counts are exact for all cells in these strata
  - Implies that the correct number of records and the correct number of records for voting-age persons is known in each cell
- For each tract and block within tract:
  - Use every zero in the published tables to eliminate rows among the 28,980 feasible micro-data images (a zero at the tract level eliminates the combination for all blocks on that tract)
  - Select the first feasible multiset of records from among those that remain such that when the reconstructed micro-data are tabulated they match every count in the selected tract and block tables
- This is standard large-scale linear equation system that can be solved by open source and commercial software
- Because of its structure, the system is massively parallel in tracts
- Blocks within tract are solved as a group

Solving the Equation System II

- Whether the problem is overdetermined (too many equations; no exact solution), exact (one unique solution), or underdetermined (too few equations; many exact solutions) depends upon the sparsity of the tables.
  - Because the tables originated from a single micro-data file (Hundred-percent Detail File, HDF), an overdetermined system implies an error in the problem set-up; there can never be more numbers in the published tables than can be created from HDF
  - When the system is exact, only one configuration (multiset) from the sample space could have produced the published tables—the reconstruction is exact
  - When the system is underdetermined there are infinitely many ways the records in the sample space could be selected to get the same publication tables
- Even when the system is underdetermined, all solutions could share some exact images
  - For example, every 2010 reconstruction has exactly the same block-level geocode and voting age values

Formal Privacy

2006: Differential Privacy

Abstract. We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f = \sum_i g(x_i)$, where x_i denotes the ith row of the database and g maps database rows to $[0, 1]$. We extend the study to general functions f, proving that privacy can be preserved by calibrating the standard deviation of the noise according to the *sensitivity* of the function f. Roughly speaking, this is the amount that any single argument to f can change its output. The new analysis shows that for several particular applications substantially less noise is needed than was previously understood to be the case.

The first step is a very clean characterization of privacy in terms of indistinguishability of transcripts. Additionally, we obtain separation results showing the increased value of interactive sanitization mechanisms over non-interactive.

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3]*

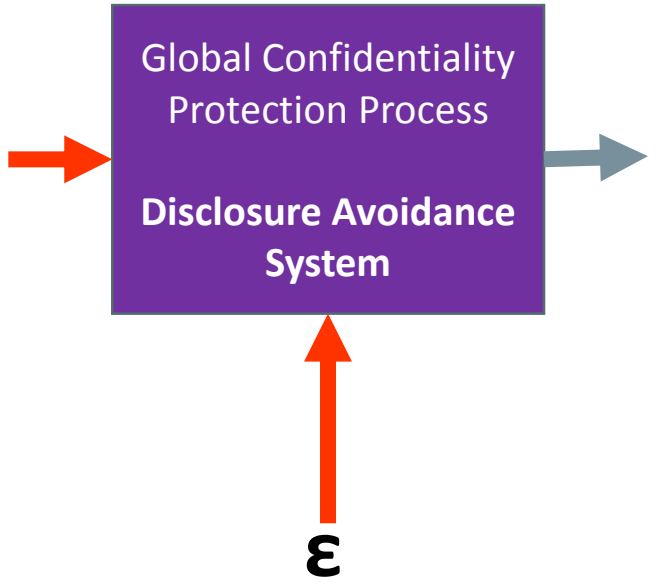
[1] Microsoft Research, Silicon Valley. {dwork,mcsherry}@microsoft.com
[2] Ben-Gurion University. kobbi@cs.bgu.ac.il
[3] Weizmann Institute of Science. adam.smith@weizmann.ac.il

Abstract. We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called *true answer* is the result of applying

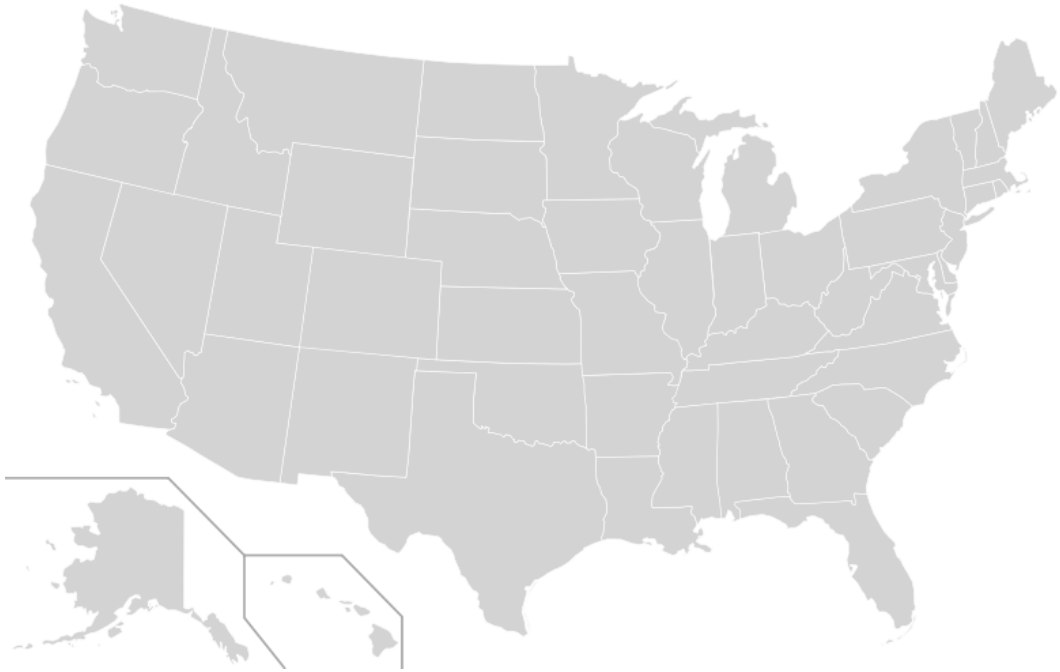
* Supported by the Louis L. and Anita M. Perlman Postdoctoral Fellowship.

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

- Advantages of noise injection with formal privacy:
  - Privacy operations are *composable*
  - Privacy guarantees are robust to post-processing
  - Provable and *tunable* privacy guarantees
  - Protects against database reconstruction attacks
  - Easy to understand
- Disadvantages:
  - Entire country must be processed at once for best accuracy
  - Every use of private data must be tallied in the *privacy-loss budget*



2020 Census of Population and Households



United States Census 2020

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

The Top-Down Algorithm

National table of US population

2 x 126 x 17 x 115

Spend ε_1 privacy-loss budget

National table with all 500,000 cells filled, structural zeros imposed with accuracy allowed by ε_1
2 x 126 x 17 x 115

Sex: Male / Female
Race + Hispanic: 126 possible values
Relationship to Householder: 17
Age: 0-114

Reconstruct individual micro-data without geography

330,000,000 records

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

State-level

State-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend ε_2 privacy-loss budget

Target state-level tables required for best accuracy for PL-94 and SF-1

Construct best-fitting individual micro-data with state geography

330,000,000 records now including state identifiers

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

County-level

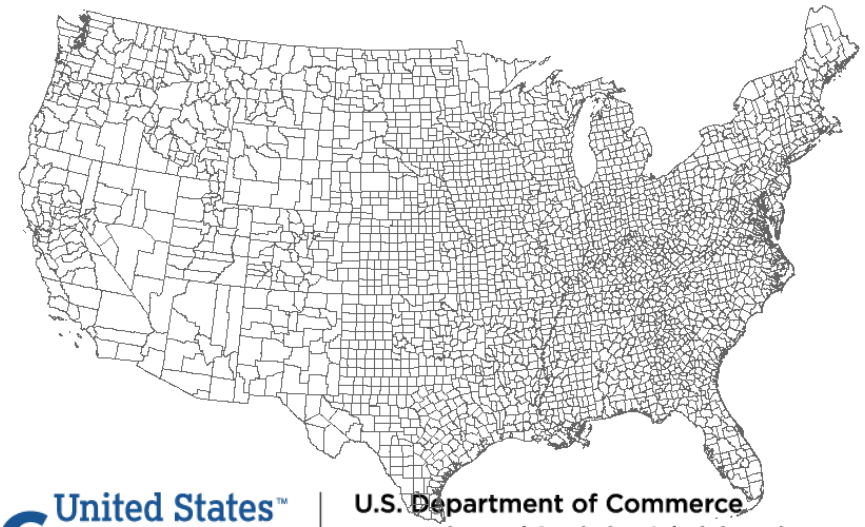
County-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend ε_3 privacy-loss budget

Target county-level tables required for best accuracy for PL-94 and SF-1

Construct best-fitting individual micro-data with state and county geography

330,000,000 records now including state and county identifiers



United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

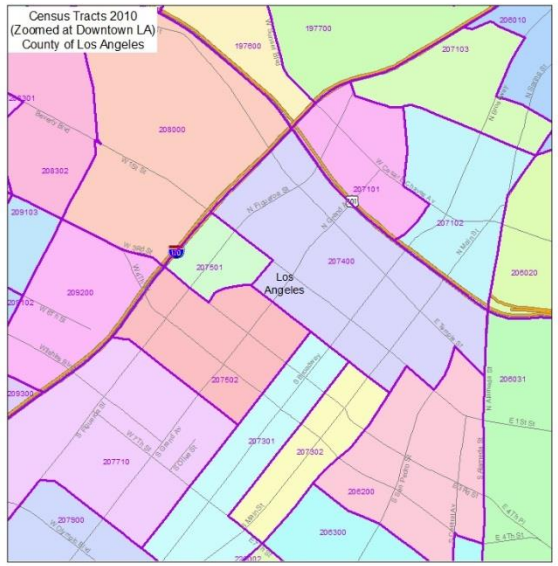
identifiers

Census tract-level

Tract-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend ε_4 privacy-loss budget

Target tract-level tables required for best accuracy for PL-94 and SF-1



Construct best-fitting individual micro-data with state, county, and tract geography

330,000,000 records now including state, county, and tract identifiers

Unit Census Bureau

'ce

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

tract identifiers

Block-level

Block-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend ε_5 privacy-loss budget

Block tract-level tables required for best accuracy for PL-94 and SF-1

Construct best-fitting individual micro-data with state, county, tract and block geography

330,000,000 records now including state, county, tract identifiers

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Tabulation micro-data

Construct best-fitting individual micro-data with state, county, tract and block geography

330,000,000 records now including state, county, tract, and block identifiers

Micro-data used for tabulating PL-94, SF-1

Tabulation micro-data

- How accurate are the tabulation micro-data?

Disclosure Avoidance Certificate

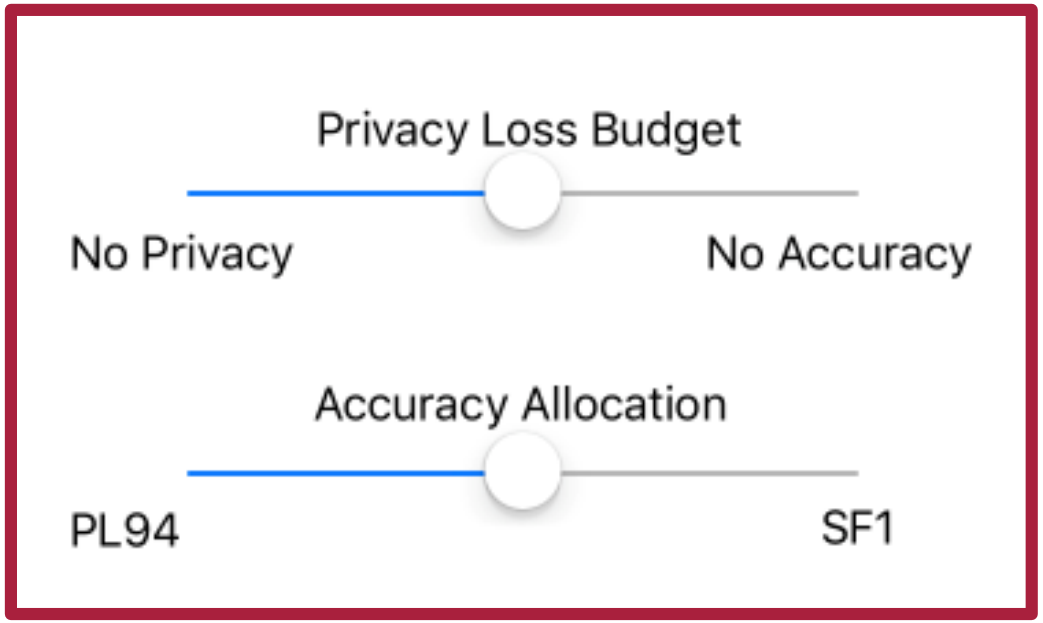
- Certifies that the disclosure avoidance system passed all tests
- Reports the accuracy of the micro-data used for tabulation
- Requires $\boldsymbol{\varepsilon}_A$

Construct best-fitting individual micro-data with state, county, tract and block geography

330,000,000 records now including state, county, tract, and block identifiers

Micro-data used for tabulating
PL-94, SF-1

Operational Decisions



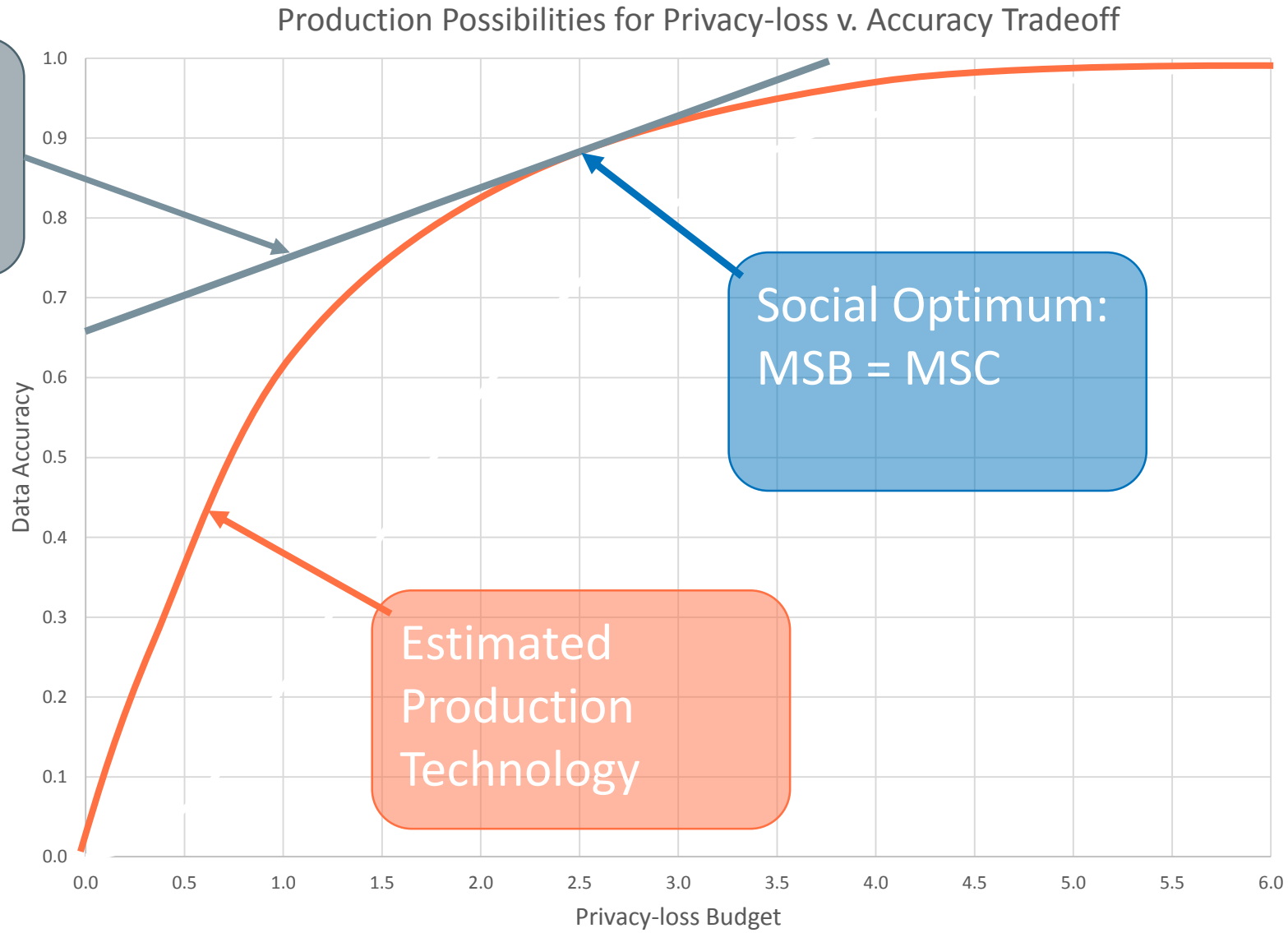
- Set total privacy loss budget: $\boldsymbol{\varepsilon}$
  - Ensure that $\boldsymbol{\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_A = \varepsilon}$
- Within each stage, allocate privacy-loss budget between:
  - PL-94
  - Parts of SF-1 not in PL-94
- These are policy levers provided by the system.
- Levers are set by the Census Bureau's Data Stewardship Executive Policy Committee

Managing the Tradeoff

How to Think about the Social Choice Problem

- The marginal social benefit is the sum of all persons' willingness-to-pay for data accuracy with increased privacy loss
- The next slide shows an example
- This is exactly the same problem being addressed by Google in RAPPOR, Apple in iOS 11, and Microsoft in Windows 10

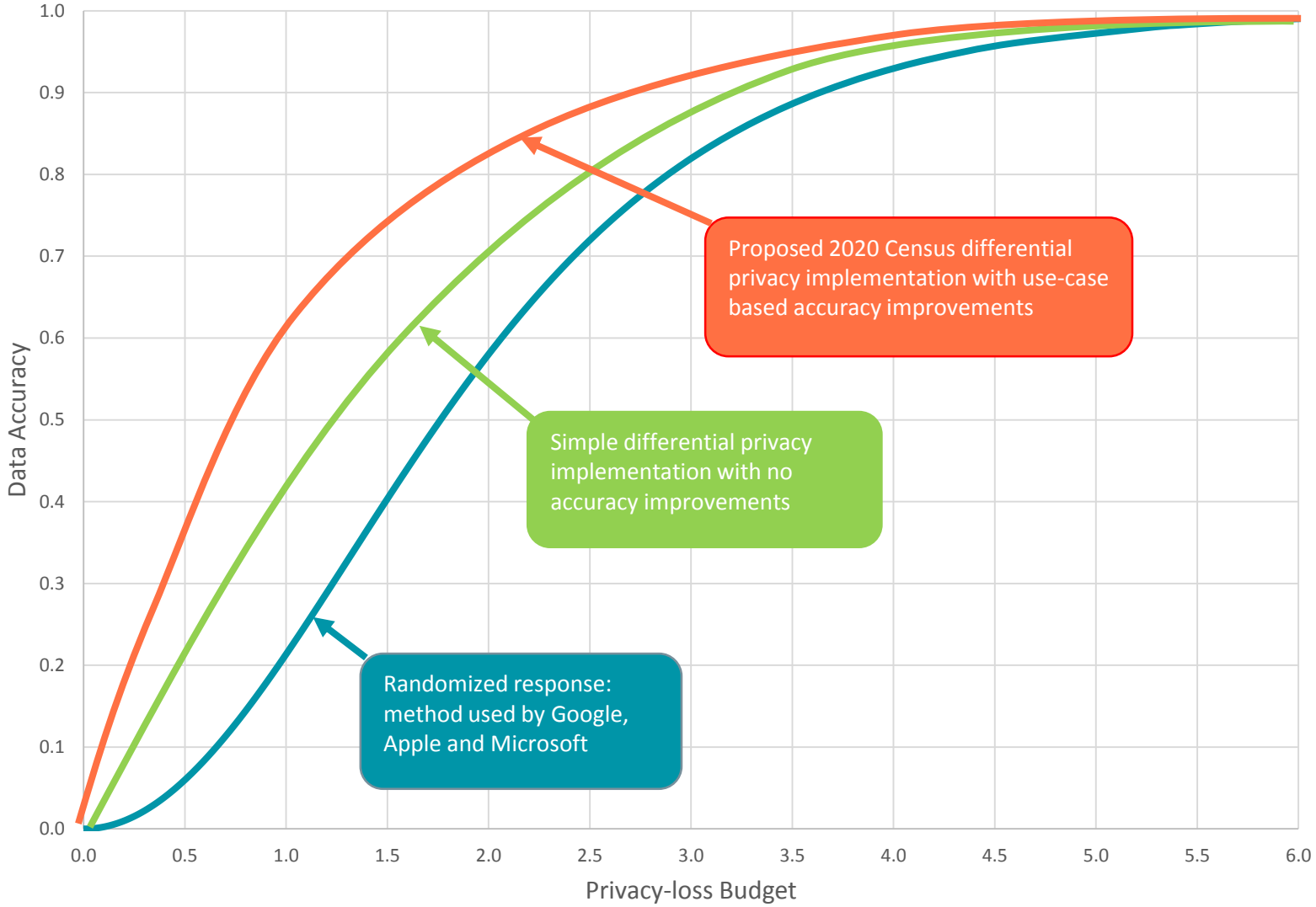
Estimated Marginal Social Benefit Curve

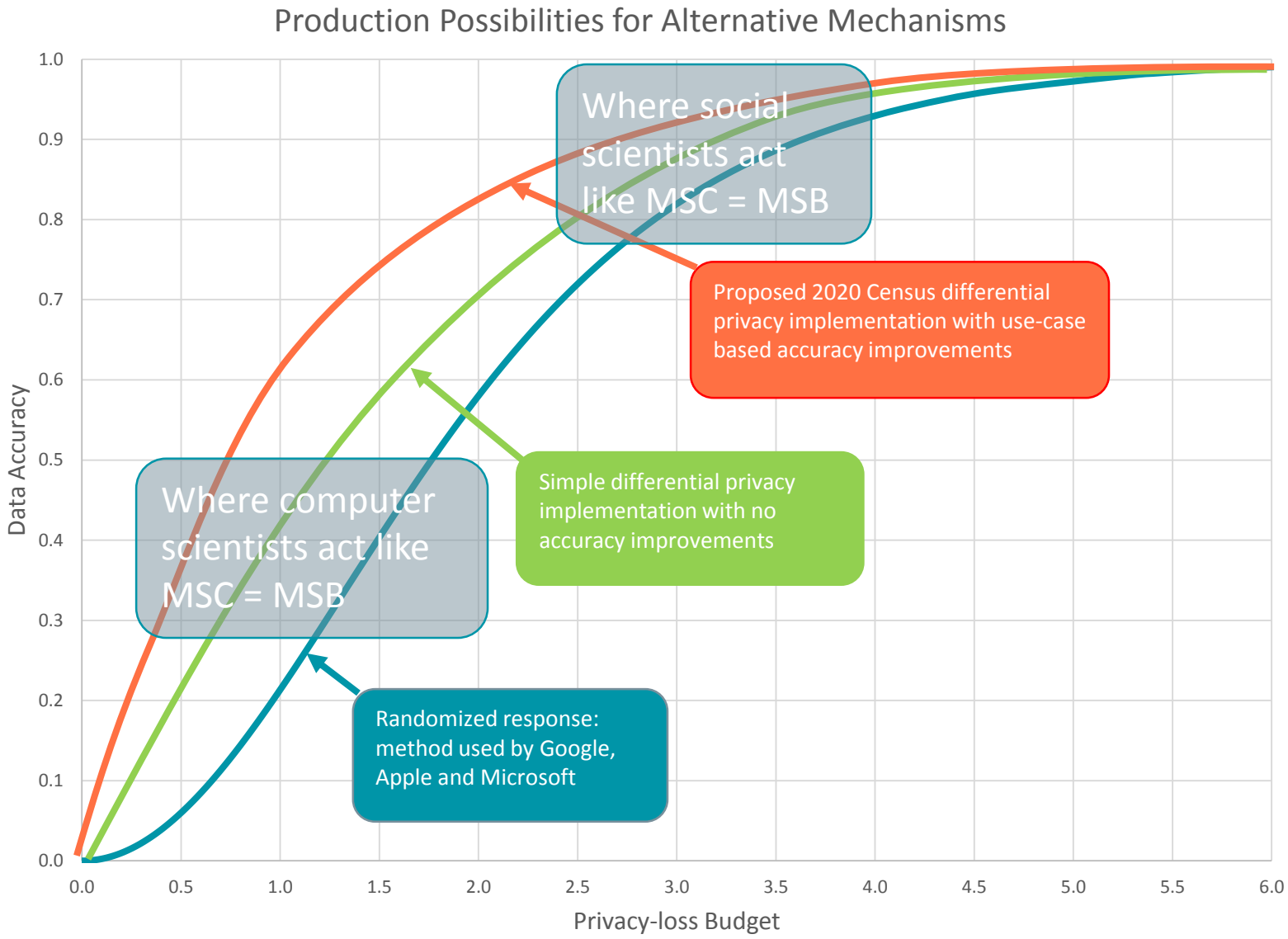


But the Choice Problem for PL94-171 Tabulations Is More Challenging

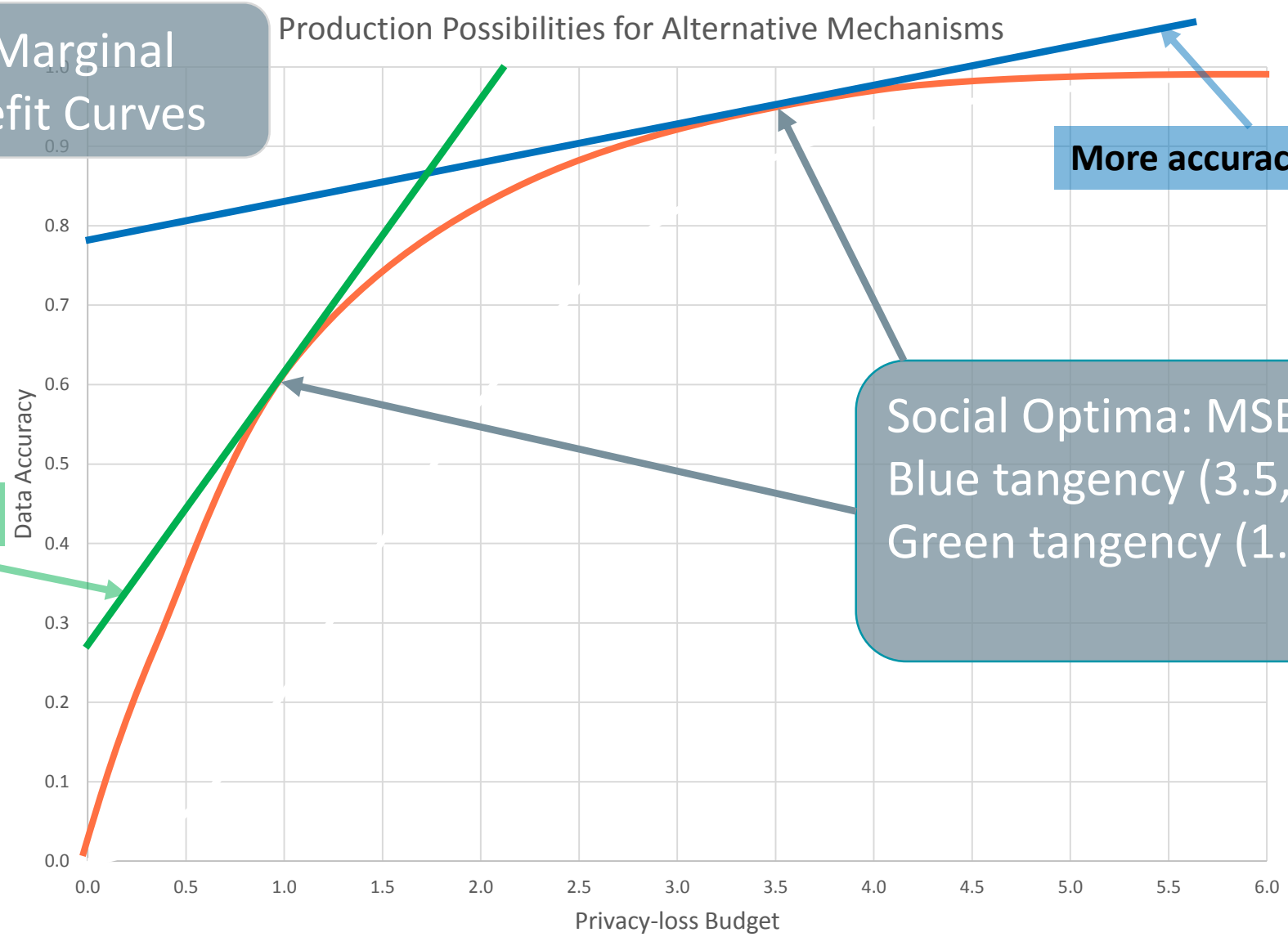
- In the redistricting application, the fitness-for-use is based on
  - Supreme Court one-person one-vote decision (All legislative districts must have approximately equal populations; there is judicially approved variation)
  - *Is statistical disclosure limitation a "statistical method" (permitted 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 and ethnicity

Production Possibilities for Alternative Mechanisms




Production Possibilities for Alternative Mechanisms
Data Accuracy
Privacy-loss Budget
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
Where social scientists act like MSC = MSB
Proposed 2020 Census differential privacy implementation with use-case based accuracy improvements
Simple differential privacy implementation with no accuracy improvements
Where computer scientists act like MSC = MSB
Randomized response: method used by Google, Apple and Microsoft

Estimated Marginal Social Benefit Curves



More accuracy favoring

More privacy favoring

Social Optima: MSB = MSC
Blue tangency (3.5, 94%)
Green tangency (1.0, 60%)

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.
- 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.
- 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, at https://digitalcommons.ilr.cornell.edu/ldi/37/
- 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.