Why Statistical Agencies Need to Take Privacy-loss Budgets Seriously, and What It Means When They Do

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

2016 FCSM Statistical Policy Seminar
The Future of Federal Statistics – Use of Multiple Data Sources, Anchored in Fundamental Principles and Practices
December 6-7, 2016

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

Acknowledgments and Disclaimer

- 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 the my own

Outline

- The database reconstruction theorem, a.k.a. the fundamental law of information reconstruction
- 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?

The Database Reconstruction Theorem

- Powerful result originally proven in Dinur and Nissim (2003) [link]
- Too many statistics published too accurately from a finite confidential database exposes the entire database with certainty
- How accurately is “too accurately”? For counts, the most common type of statistical summary, the cumulative noise in the published statistics must be of the magnitude $\sqrt{N}$

Database Reconstruction II

- This work led quickly to the first important results in privacy-preserving data analysis with formal (mathematical) guarantees:
  - Dwork, McSherry, Nissim, and Smith (2006) [link]
  - Dwork (2006) [link]
- This particular privacy-preserving system is known as "differential privacy," about which more later

Database Reconstruction III

- Dwork, McSherry and Talwar (2007) [link]

  "In the context of privacy-preserving datamining our results say that any privacy mechanism, interactive or non-interactive, providing reasonably accurate answers to a 0.761 fraction of randomly generated weighted subset sum queries, and arbitrary answers on the remaining 0.239 fraction, is blatantly non-private."

- Muthukrishnan and Nikolov (2012) [link]

  Improves the Dinur-Nissim bound

- Kasiviswanathan, Rudelson and Smith (2013) [link]

  Extends the database reconstruction model to include linear and logistic regression, M-estimators, classifiers, decision trees, contingency tables

- Dwork, Smith, Steinke, Ullman, and Vadhan (2015) [link]

  Shows that exact knowledge of even a single exact case can compromise even noisy publications from genome-wide association studies, other Bernoulli data, and multivariate normal data, and not just for that case

- Cynthia Dwork has recently labeled this collection of results "The Fundamental Law of Information Recovery" (Dwork and Roth, 2014 [link], Dwork, undated [link])

Historical Note

- The U.S. Census Bureau was the first organization anywhere in the world to use a formally private confidentiality protection system in production
  - OnTheMap (residential side)
- Machanavajjhala, Kifer, Abowd, Gehrke, and Vilhuber (2008) [link] implements a variant of differential privacy known as "probabilistic differential privacy," which is very similar to what is now called (ε, δ)-differential privacy

What is a Privacy-loss Budget?

- Not a dollar budget, but it works the same way
- Cumulative summary of the aggregate risk of partial database reconstruction given all of the statistics published to date
- In privacy-preserving data analysis (also called formal privacy modeling) the privacy-loss budget is a worst-case limit to the inferential disclosure of any identity or item in the confidential database
- In the leading example of privacy-preserving data analysis, "differential privacy," this worst case is over all possible databases with the same schema (sample space to statisticians) and all individuals and items

Why Use Worst-case Protection?

- The interpretation of a privacy-loss budget as “worst-case” protection is a reasonable application of “equal protection under the law”
- It protects a hypothetical person who meets the definition of being in the protected population
- The hypothetical person is anyone who is in the universe of the database schema (sample space)—anyone who might have been selected for the census or survey
- “Average-case” protection does not have this property—one can identify who is advantaged or disadvantaged *a priori*

Respecting a Privacy-loss Budget

- Respecting a privacy-loss budget means that the entire collection of statistics released from a single confidential database can *never* permit a database reconstruction that is more accurate on any data item or individual than the limit given by the budget
- The protection into the indefinite future applies to all potential future attackers and all potential future information of the form specified in the privacy model
- For “differential privacy” this guarantee is over all future attackers and any database with the same schema (sample space)

Current Context

- Don’t current confidentiality laws require data stewards to respect a privacy-loss budget, at least implicitly?
- Unclear
- Current confidentiality laws prohibit the stewards from publishing exactly identifiable items or individuals, but they are silent on the subject of limiting inferential disclosure: 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
  - See Dwork and Naor (2008) [link] for the “impossibility theorem”
  - See Kifer and Machanavajjhala (2011) [link] for the “no free lunch theorem”

This Is Not a New Problem

- Ancient civilizations knew that the ratio of the area of a circle to its diameter was constant, but they 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: you can't square a circle, no rational number is exact for this problem
- But even after the discovery of irrational numbers, legal documents continued to use crude approximations:
  - Indiana state legislature (1897), other examples
- It takes human societies 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 (asking a survey respondent one of two questions at random: one sensitive (SQ), one innocuous; the innocuous question can also be randomized) is provably privacy-loss protective
- Privacy loss is bounded by the maximum Bayes factor designed into the randomized response protocol

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

- In privacy-preserving data analysis this bound is stated as the logarithm of the maximum Bayes factor
- If the sensitive question is asked with probability ½ and the 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 (for all possible input databases and all future attacks of any form)

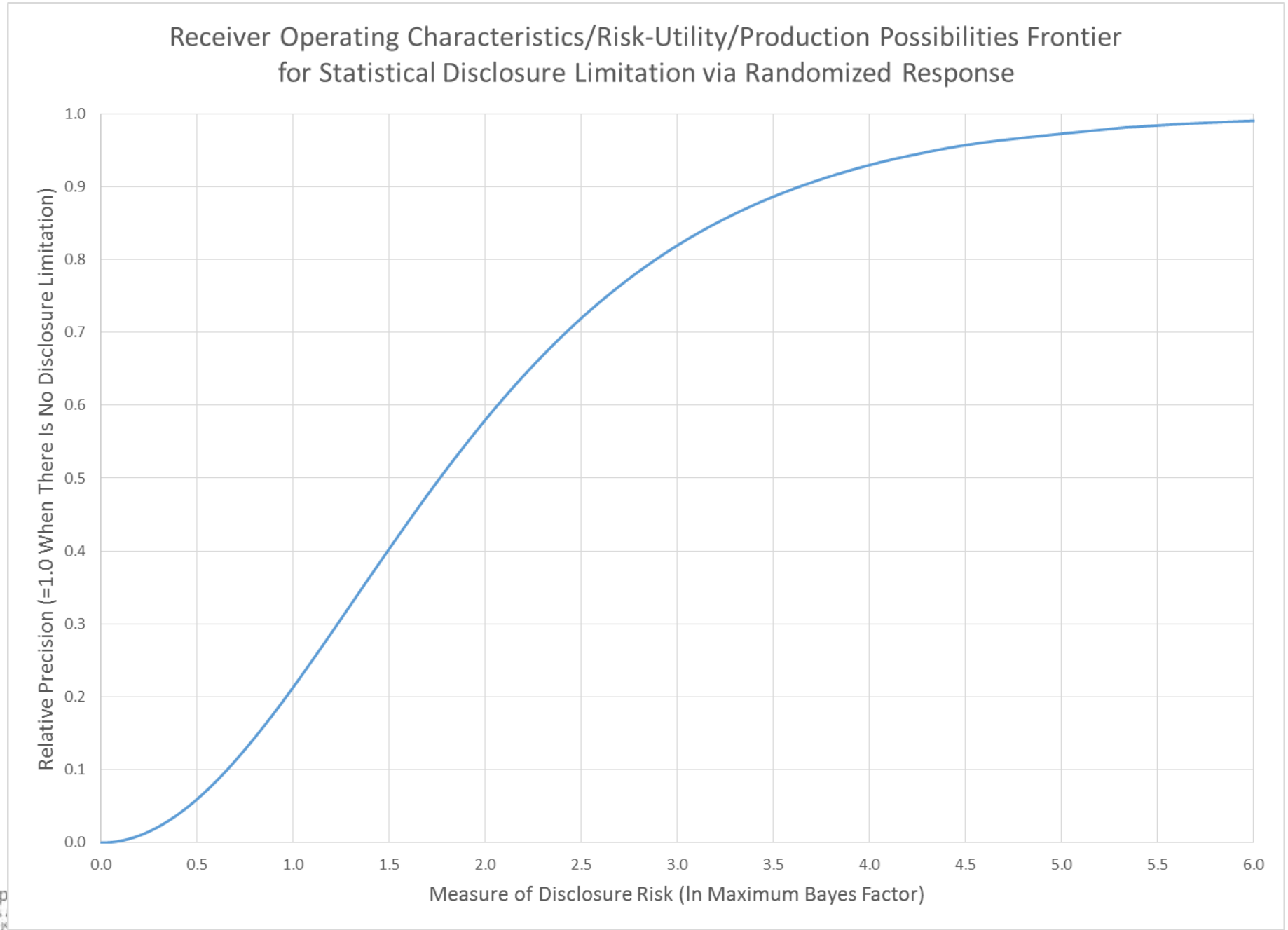
Sources: Warner (1965) [link] and Greenberg, Abdel-Latif, Simmons, and Horvitz (1969) [link]. SDL uses: Fienberg and Steele (1998) [link], Du and Zhan (2003) [link] and Erlingsson, Vasyl and Korolova (2014) [link].

What Happens to Data Quality?

- The sampling precision (inverse of the sampling variance) associated with the randomized-response estimator is proportional to the sampling precision when all respondents are asked the sensitive question

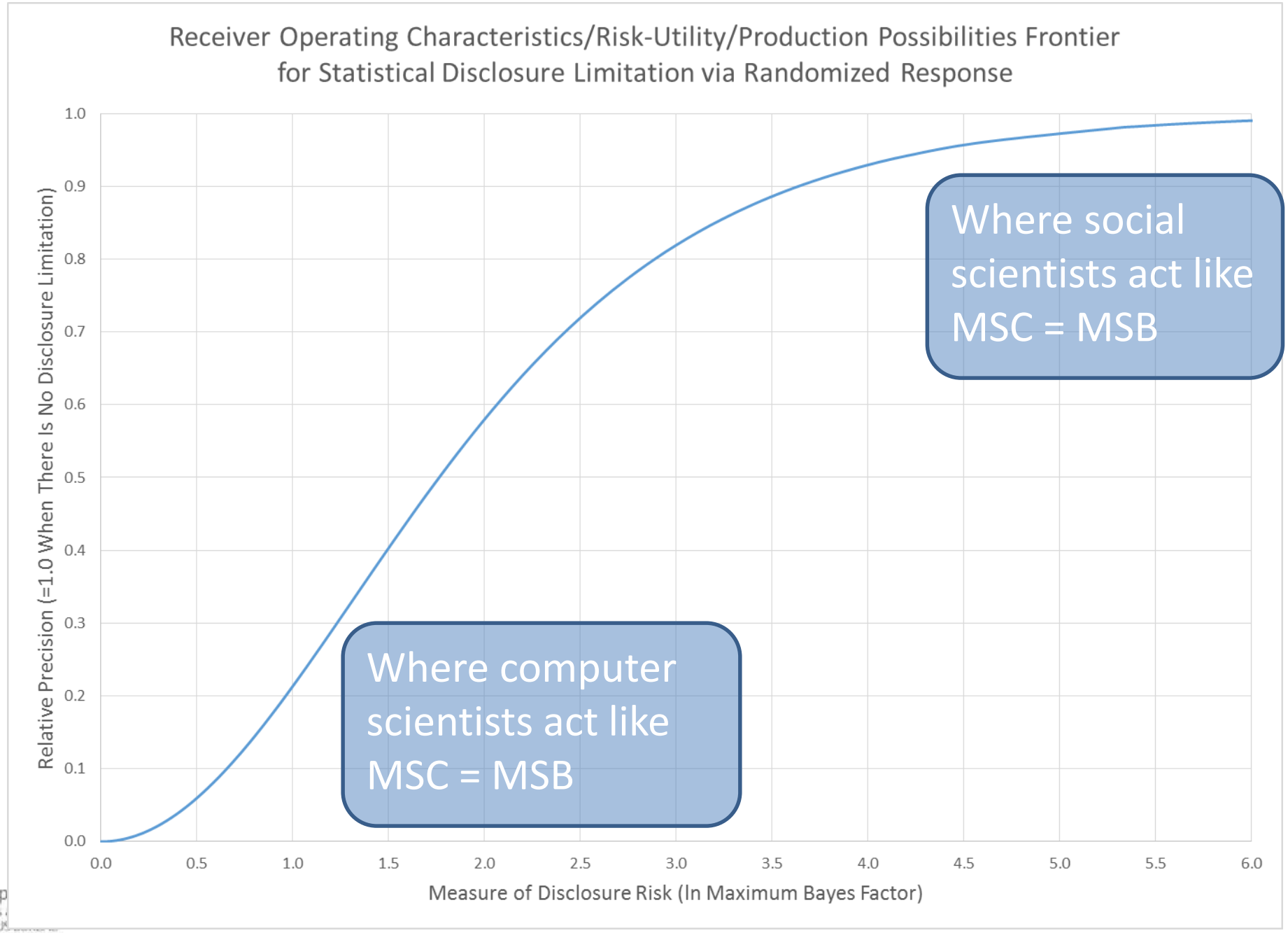
$$Rel.Precision = \frac{\{Pr[Ask\ Sensitive\ Q]\}^2 \frac{n}{\theta(1-\theta)}}{\frac{n}{\theta(1-\theta)}} = \left\{\frac{1}{2}\right\}^2 = 0.25$$

- Under the same conditions in which the privacy loss is ln 3, the relative sampling precision is 25% of the most accurate estimator
- The graph shows this trade-off from privacy-loss expenditures of zero to 6 (ε-differential privacy, x-axis)
- Data quality is measured by the relative precision (y-axis)


Receiver Operating Characteristics/Risk-Utility/Production Possibilities Frontier
for Statistical Disclosure Limitation via Randomized Response
Relative Precision (=1.0 When There Is No Disclosure Limitation)
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
Measure of Disclosure Risk (ln Maximum Bayes Factor)

Disclosure Limitation is Technology

- Properly designed methods operate with a privacy-loss budget constraint
- The price of increasing data quality (more accurate publications) in terms of increased privacy loss 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


Receiver Operating Characteristics/Risk-Utility/Production Possibilities Frontier
for Statistical Disclosure Limitation via Randomized Response
Relative Precision (=1.0 When There Is No Disclosure Limitation)
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
Measure of Disclosure Risk (In Maximum Bayes Factor)
Where social scientists act like MSC = MSB
Where computer scientists act like MSC = MSB

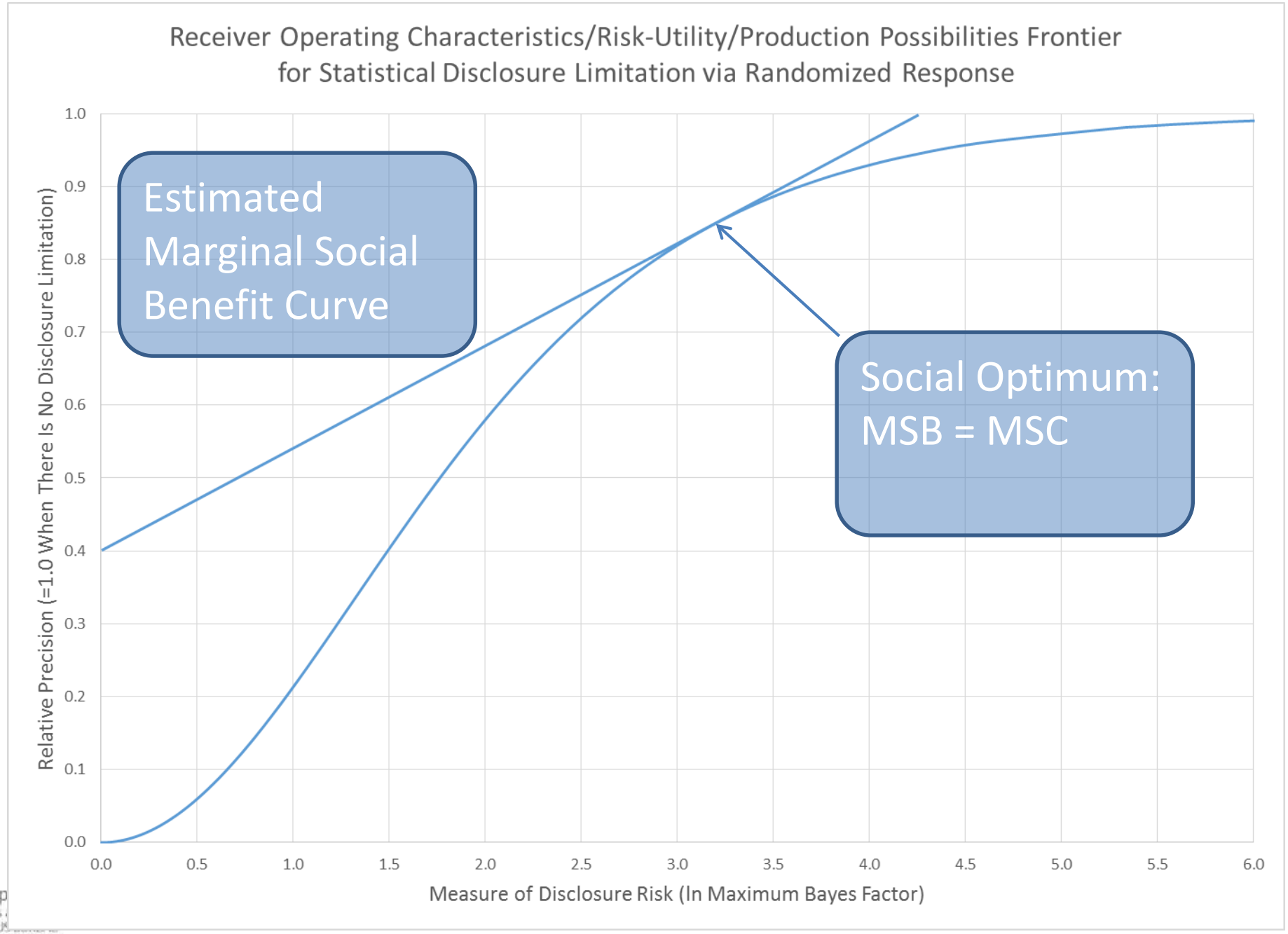
Some Examples

- Dwork (2008): “The parameter ε in Definition 1 is public. The choice of ε is essentially a social question and is beyond the scope of this paper. That said, we tend to think of ε as, say, 0.01, 0.1, or in some cases, ln 2 or ln 3. If the probability that some bad event will occur is very small, it might be tolerable to increase it by such factors as 2 or 3, while if the probability is already felt to be close to unacceptable, then an increase by a factor of $e^{0.01} \approx 1.01$ might be tolerable, while an increase of e, or even only $e^{0.1}$, would be intolerable.” [link, p. 3]
- Dwork (2011): “The parameter ε 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, $\varepsilon = 8.9$, which was required to produce tract-level estimates with acceptable accuracy

How to Think about the Social Choice Problem

- The marginal social benefit from increased data quality is the sum of all citizens' willingness-to-pay for data quality with increased privacy loss
- Can be estimated from survey data
- The next slide shows how to use the estimate from survey data to select an optimal (data quality, privacy loss) pair

See Abowd and Schmutte (2015) [link].


Receiver Operating Characteristics/Risk-Utility/Production Possibilities Frontier
for Statistical Disclosure Limitation via Randomized Response
Relative Precision (=1.0 When There Is No Disclosure Limitation)
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
Measure of Disclosure Risk (ln Maximum Bayes Factor)
Estimated Marginal Social Benefit Curve
Social Optimum: MSB = MSC

How to Prove That a Privacy-loss Budget Was Respected

- You can't respect a privacy-loss budget without quantifying the expenditure of each publication
- The collection of the algorithms taken altogether must satisfy the privacy loss budget
- Requires methods that have known aggregation properties (called "composition theorems" in algorithmic design)

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

- The information environment is changing much faster than it was when most current disclosure limitation methods were invented
- It may no longer be reasonable to assert that a product is empirically safe given best-practice disclosure limitation prior to its release
- But, in the randomized response example, the privacy loss of ln 3 never grows regardless of any future data publications from external sources, this is what formal privacy-loss budgets guarantee
- Resistance to future attacks is a design property of the methods, not an empirical property of the application

The Silver Lining

- The American Statistical Association has recently released a statement on statistical significance and p-values [link]
- This was widely interpreted as a call for more nuanced use of the p-value and statements about statistical significance
- Under very general conditions, data analysis conducted using privacy-preserving methods can control the false discovery rate and reduce inferential errors due to multiple comparisons
  - Examples: Erlingsson, Vasyl and Korolova (2014) [link]; Dwork et al. (2015) [link]

A Long Row to Hoe

- It is going to take a concerted research and engineering effort to bring disclosure limitation into the 21st century
- But the scientific integrity of agency enterprises requires that we tackle this challenge
- The 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

Thank you.

john.maron.abowd@census.gov