12-7-2016

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

John M. Abowd
Cornell University, John.Abowd@cornell.edu

Follow this and additional works at: http://digitalcommons.ilr.cornell.edu/ldi
Thank you for downloading an article from DigitalCommons@ILR.
Support this valuable resource today!
Why Statistical Agencies Need to Take Privacy-loss Budgets Seriously, and What It Means When They Do

Comments
To appear on fcsm.sites.usa.gov, as presented to the 2016 FCSM Statistical Policy Seminar.

Abowd acknowledges partial support through NSF Grants 1131848 (NCRN) and 1012593 (TC:Large), support by the Alfred P. Sloan Foundation, and from the U.S. Census Bureau.

At the time this talk was given, Abowd was Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau. The opinions expressed in this talk are his own.

This article is available at DigitalCommons@ILR: http://digitalcommons.ilr.cornell.edu/ldi/32
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
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 recovery
- 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 from Dinur and Nissim (2003) [link]
- Too many statistics published too accurately from a confidential database exposes the entire database with certainty
- How accurately is “too accurately”?
  - Cumulative noise must be of the order $\sqrt{N}$
Database Reconstruction II

- Led quickly to “differential privacy”:
  - Dwork, McSherry, Nissim, and Smith (2006) [link]
  - Dwork (2006) [link]
- Leading formal privacy model
Database Reconstruction III

  - Dwork and Roth, 2014 [link]
  - Dwork, undated [link]
- Includes extensions found in
  - Dwork, McSherry and Talwar (2007) [link]
  - Muthukrishnan and Nikolov (2012) [link]
  - Kasiviswanathan, Rudelson and Smith (2013) [link]
  - Dwork, Smith, Steinke, Ullman, and Vadhan (2015) [link]
Historical Note

- The U.S. Census Bureau: first organization in the world to use a formally private confidentiality protection system in production
  - OnTheMap (residential side)
- Machanavajjhala, Kifer, Abowd, Gehrke, and Vilhuber (2008) [link]
What is a Privacy-loss Budget?

- Not a dollar budget, but works the same way
- Constrains aggregate risk of partial database reconstruction given all published statistics
- Worst-case limit to the inferential disclosure of any identity or item
- In differential privacy, worst case is over all possible databases with the same schema for all individuals and items
Why Use Worst-case Protection?

- “Worst case” is “equal protection under the law”
  - Protects every person in the population the same way
  - Anyone who might have been selected for the census or survey, whether in the database or not
- “Average-case” protection does not
  - Can identify who is advantaged or disadvantaged \textit{a priori}
Respecting a Privacy-loss Budget

- All released statistics can *never* permit a database reconstruction more accurate than the budget
- Protection into the indefinite future
- For differential privacy, guarantee is over all future attackers and any database with the same schema
Current Context

- Don’t current confidentiality laws require data stewards to respect a privacy-loss budget, at least implicitly?
- Unclear
- Law are silent on limitations of 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
  - Dwork and Naor (2008) [link]: impossibility theorem
  - Kifer and Machanavajjhala (2011) [link]: no free lunch theorem
This Is Not a New Problem

- Ratio of the circumference of a circle to its diameter is constant
- Ancients didn’t understand irrational numbers:
  - Babylonians: $\pi = 3 \frac{1}{8}$
  - Egyptians: $\pi = 4 \times (\frac{8}{9})^2$
  - Israelites: $\pi = 3$ [Talmud legislated value]
  - Hindu: $\pi = \frac{62,832}{20,000} = 3.1416$
  - Euclid: no rational number is exact for this problem
  - Archimedes: sequences can approximate $\pi$ with increasing accuracy
- But legal documents continued to use crude approximations
- Takes 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 is provably privacy-loss protective
- Privacy loss bounded by the maximum Bayes factor

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

- Bound is the logarithm of the maximum Bayes factor
- If
  - Sensitive question asked with probability \( \frac{1}{2} \)
  - And innocuous question is “yes” with probability \( \frac{1}{2} \)
  - Then the maximum Bayes factor is 3, and \( \ln 3 = 1.1 \)
- The privacy-loss expenditure (\( \epsilon \)-differential privacy) is 1.1
What Happens to Data Quality?

- Use relative sampling precision

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

- If

  - Privacy loss is \(\ln 3\)
  - Then, relative sampling precision is 25% of the most accurate estimator
Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics for Statistical Disclosure Limitation via Randomized Response
Disclosure Limitation is Technology

- The price of increasing data quality (public “good”) in terms of increased privacy loss (public “bad”) 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
Data Quality (Relative Precision=1.0 When There Is No Privacy) -- Public "Good"

Privacy Loss Budget (ln Maximum Bayes Factor) -- Public "Bad"

Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics for Statistical Disclosure Limitation via Randomized Response

Where computer scientists act like MSC = MSB

Where social scientists act like MSC = MSB
Some Examples

- Dwork (2008): “The parameter e in Definition 1 is public. The choice of e is essentially a social question and is beyond the scope of this paper.” [link, p. 3]

- Dwork (2011): “The parameter e is public, and its selection is a social question. We tend to think of e as, say, 0.01, 0.1, or in some cases, ln 2 or ln 3.” [link, p. 91]

- In OnTheMap, e = 8.9, was required to produce tract-level estimates with acceptable accuracy
How to Think about the Social Choice Problem

- The marginal social benefit 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

See Abowd and Schmutte (2015) [link].
Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics for Statistical Disclosure Limitation via Randomized Response

Data Quality (Relative Precision=1.0 When There Is No Privacy)--Public "Good"

Privacy-Privacy Loss Budget (ln Maximum Bayes Factor)--Public "Bad"

Estimated Marginal Social Benefit Curve

Social Optimum: MSB = MSC
How to Prove That a Privacy-loss Budget Was Respected

- Must quantify the privacy-loss expenditure of each publication
- The collection of the algorithms taken altogether must satisfy the privacy-loss budget
- Requires methods that compose
How to Prove That the Algorithms are Resistant to All Future Attacks

- Information environment is changing much faster than before
- *It may no longer be reasonable to assert that a product is empirically safe given best-practice disclosure limitation prior to its release*
- Formal privacy models replace empirical assessment with designed protection
- Resistance to all future attacks is a property of the design
The Silver Lining

- American Statistical Association on p-values [link]
- Call for more nuanced use
- Data analysis conducted using privacy-preserving methods:
  - Control the false discovery rate
  - Reduce inferential errors due to multiple comparisons
  - Examples: Erlingsson, Vasyl and Korolova (2014) [link]; Dwork et al. (2015) [link]; Apple (2016) [link]
A Long Row to Hoe

- Concerted research and engineering effort needed to bring disclosure limitation into the 21st century
- Scientific integrity requires that we tackle this challenge
- 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