An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices

John M. Abowd[1,3] Ian M. Schmutte[2]

[1]Associate Director for Research and Methodology and Chief Scientist U.S. Census Bureau

[2]University of Georgia

[3]Cornell University

SOLE 2018
Toronto, ON
May 4, 2018

Problem

Data custodians trade off

- Providing detailed and accurate statistics
- Protecting privacy and confidentiality

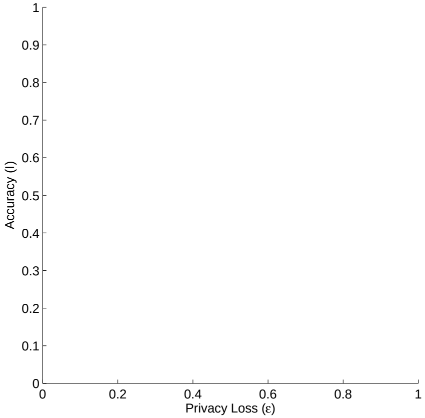
What is the optimal tradeoff, given that the data have already been collected?

Economic Approach

1. Finite resource: Information in an existing database
2. Competing uses:
   - Statistical accuracy, versus
   - Data privacy
3. An optimal allocation should equate
   - Marginal Rate of Transformation
   - Willingness to Pay (Marginal Rate of Substitution)
4. Accuracy and privacy are public goods

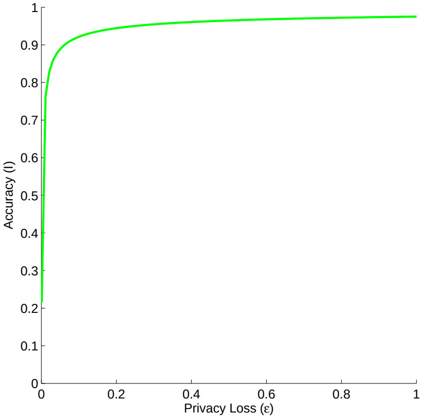
Social Welfare Maximization

Social planner's problem: Maximize welfare subject to the PPF



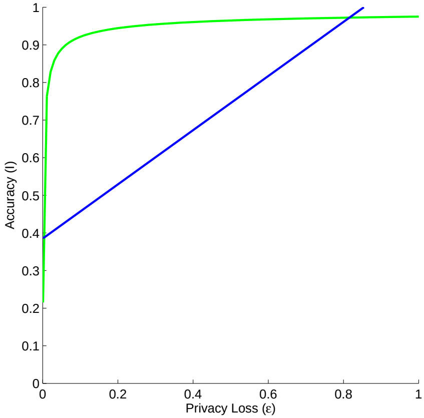
Social Welfare Maximization

Social planner's problem: Maximize welfare subject to the PPF



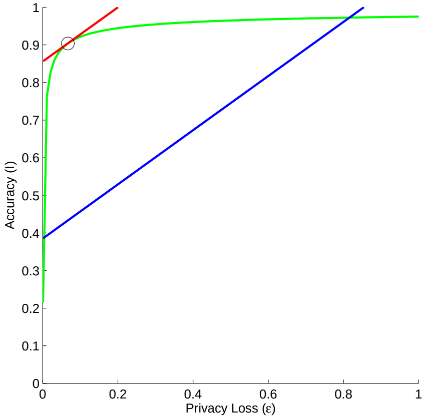
Social Welfare Maximization

Social planner's problem: Maximize welfare subject to the PPF



Social Welfare Maximization

Social planner's problem: Maximize welfare subject to the PPF

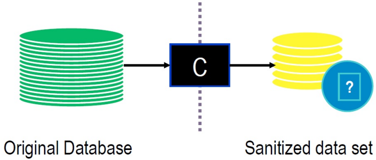


Motivation

- *Database Reconstruction Theorem* and *Fundamental Law of Information Recovery*
  - [Dinur and Nissm (2003); Dwork, McSherry, Talwar (2007)]
  - Publication of “too many” statistics with “too much” accuracy is *blatantly non-private*

Model Overview

- Data Custodian
- Existing database, D
- Desired statistics, or queries, Q
- Publication mechanism: $M(D, Q)$



Differential Privacy and Inferential Disclosure

Mechanism M is ε-*differentially private* if

$$\ln\left(\frac{\Pr\left[M(x,Q)\in B\,|x,Q\right]}{\Pr\left[M(x',Q)\in B\,|x',Q\right]}\right)\leq\varepsilon$$

[Dwork, McSherry, Nissim and Smith (2006)]

Properties

- *Data reconstruction:* ε bounds change in output from changing input
- *Privacy loss:* ε bounds "worst-case" update about x
- *Composes:* Losses due to multiple uses of the same data are "added up"
- *Future Proof*: Guarantees independent of outside knowledge
- *Public:* Mechanism and parameters can be published [*SDL-aware analysis*]

Application to Title I

Setting

- Title I funds appropriated by Congress to needy school districts
- DOE allocates to district ℓ using

$$A_\ell = E_\ell \times C_\ell,$$

  - A_ℓ is the *authorization amount*
  - E_ℓ is the *eligibility count*
  - C_ℓ is the *adjusted per-pupil expenditure*
- Census publishes $\widehat{E}_\ell$
- **Target Allocation:** $X = \sum_{\ell=1}^{L} E_\ell \times C_\ell$
- **Actual Allocation:** $\widehat{X} = \sum_{\ell=1}^{L} \widehat{E}_\ell \times C_\ell$

Publication Mechanism

- **Database:** Households with indicator for Title I eligibility and district geocode
- **Queries:** Count of Title I households by district (E_ℓ)
- **Mechanism:** Laplace Mechanism (Matrix Mechanism)
  - Publish $\widehat{E}_\ell = E_\ell + e_\ell$
  - e_ℓ is Laplace noise with scale parameter ε^{-1}
  - Satisfies $\varepsilon-$differential privacy
  - Accuracy:

$$I = -\mathbb{E}\left[\sum_{\ell=1}^{L}\left(\widehat{E}_\ell - E_\ell\right)^2\right] = -\frac{2L}{\varepsilon^2}$$

Social Welfare Function

$$SWF = \phi \sum_i v_i^{Info}(\varepsilon) + (1-\phi) v^{Data}(I),$$

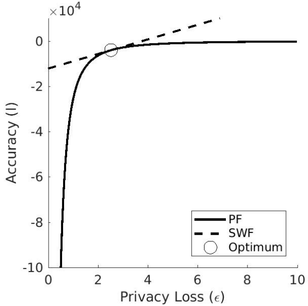
- Weight, $0 \leq \phi \leq 1$, on privacy preferences
- **Information Utility:** $v_i^{Info}(\varepsilon) = -k_i \varepsilon$ [Ghosh and Roth (2015)]
- **Data Utility:** $v^{Data}(I)$
  - Linear-quadratic in aggregate misallocation: $W = (\widehat{X} - X) = \sum_{\ell=1}^{L} C_\ell \left[\widehat{E}_\ell - E_\ell\right]$
  - $v^{Data}(I) = I \sum_{\ell=1}^{L} \frac{C_\ell^2}{L}$

Calibration

$$WTA \equiv \frac{dI}{d\varepsilon} = \left(\frac{\phi}{1-\phi}\right) N \frac{\bar{k}}{\bar{C^2}},$$

- $L = 13,000$ public school districts
- $N = 46$ million school-age children
- average squared spending, $\bar{C^2} \approx 20$ million
- $\bar{k} = \$1,400$ (avg. cost of identity theft)
- Setting $WTA = MRT$

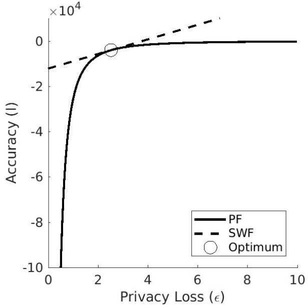
$$\varepsilon = 2.52 \times \left(\frac{\phi}{1-\phi}\right)^{-\frac{1}{3}}$$



Calibration

$$\eta = \frac{\phi}{1-\phi}$$

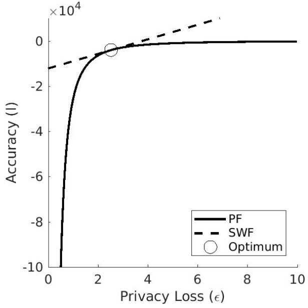
- $\eta = 1$



Calibration

$$\eta = \frac{\phi}{1 - \phi}$$

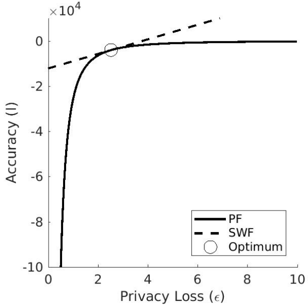
- $\eta = 1$
  - $\varepsilon^* = 2.52$
  - *RMSE* : $2,509 (70 cents per student)



Calibration

$$\eta = \frac{\phi}{1 - \phi}$$

- $\eta = 1$
  - $\varepsilon^* = 2.52$
  - $RMSE$: $2,509 (70 cents per student)
- $\eta = \frac{N}{POP - N} \approx 0.15$

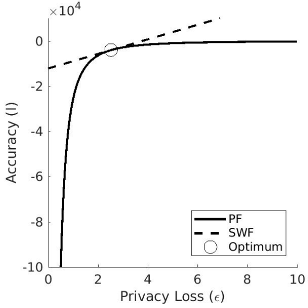


Calibration

$$\eta = \frac{\phi}{1-\phi}$$

- $\eta = 1$
  - $\varepsilon^* = 2.52$
  - $RMSE$: \$2,509 (70 cents per student)

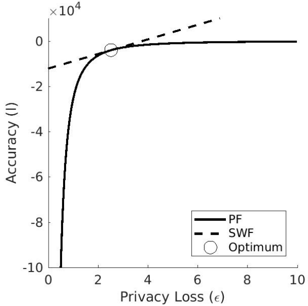
- $\eta = \frac{N}{POP-N} \approx 0.15$
  - $\varepsilon^{**} = 4.74$
  - $RMSE$: \$1,334 (38 cents per student)



Calibration

$$\eta = \frac{\phi}{1-\phi}$$

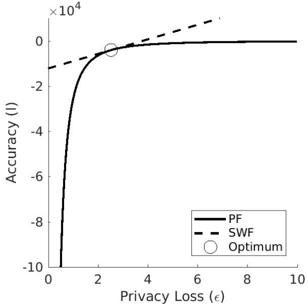
- $\eta = 1$
  - $\varepsilon^* = 2.52$
  - $RMSE$: \$2,509 (70 cents per student)
- $\eta = \frac{N}{POP-N} \approx 0.15$
  - $\varepsilon^{**} = 4.74$
  - $RMSE$: \$1,334 (38 cents per student)
- Privacy advocates urge $\varepsilon << 1$



Calibration

$$\eta = \frac{\phi}{1-\phi}$$

- $\eta = 1$
  - $\varepsilon^* = 2.52$
  - *RMSE* : $2,509 (70 cents per student)
- $\eta = \frac{N}{POP-N} \approx 0.15$
  - $\varepsilon^{**} = 4.74$
  - *RMSE* : $1,334 (38 cents per student)
- Privacy advocates urge $\varepsilon << 1$
  - Fix $\varepsilon = 0.1$
  - *RMSE* :$63,000 ($18 per student)



Future Work

- Better Models
  - Evaluating technology in real-world use cases
  - Demand for privacy
  - Demand for accuracy
- Better data
  - Census Bureau survey on privacy and accuracy attitudes
  - Experimental measures of preferences

References

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis, *Proceedings of the Third conference on Theory of Cryptography*, TCC'06, Springer-Verlag, Berlin, Heidelberg, pp. 265–284.

Dwork, C. and Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy, *Journal of Privacy and Confidentiality* **2**(1): 93–107.
URL: *http://repository.cmu.edu/jpc/vol2/iss1/8/*

Dwork, C. and Nissim, K. (2004). Privacy-preserving datamining on vertically partitioned databases, *Proceedings of Advances in Cryptology (CRYPTO)* **3152**: 528–544.
URL: *http://www.iacr.org/cryptodb/archive/2004/CRYPTO/1267/1267.pdf*

Gehrke, J., Lui, E. and Pass, R. (2011). Towards privacy for social networks: A zero-knowledge based definition of privacy, *Theory of Cryptography Conference*, Springer, pp. 432–449.

Ghosh, A. and Roth, A. (2015). Selling privacy at auction, *Games and Economic Behavior* **91**: 334–346.

References II

Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy, *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, ACM Digital Library, New York, NY, USA, pp. 193–204.
URL: *http://doi.acm.org/10.1145/1989323.1989345*