Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods

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

[1]Cornell University

[2]University of Georgia

[3]Center for Economic Studies, U.S. Census Bureau

NSF-Census Research Network Spring Meeting
National Academies of Science, Washington, DC
May 8, 2015

Problem

Data curators trade off

- Providing detailed and accurate statistics
- Protecting privacy and confidentiality

What is the optimal tradeoff, given 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. The optimal allocation should equate
   - Marginal Rate of Transformation
   - Willingness to Pay (Marginal Rate of Substitution)
4. Accuracy and privacy are public goods

Data Production Technology

Private Multiplicative Weights (PMW) (Hardt and Rothblum 2010) the custodian chooses

1. The set $\mathcal{Q} \subseteq \mathcal{F}^k$ of allowable queries
2. The number of queries to allow ($t = 1, \ldots, k$)
3. Privacy parameters (ε and δ)
4. Accuracy parameters (α and β)

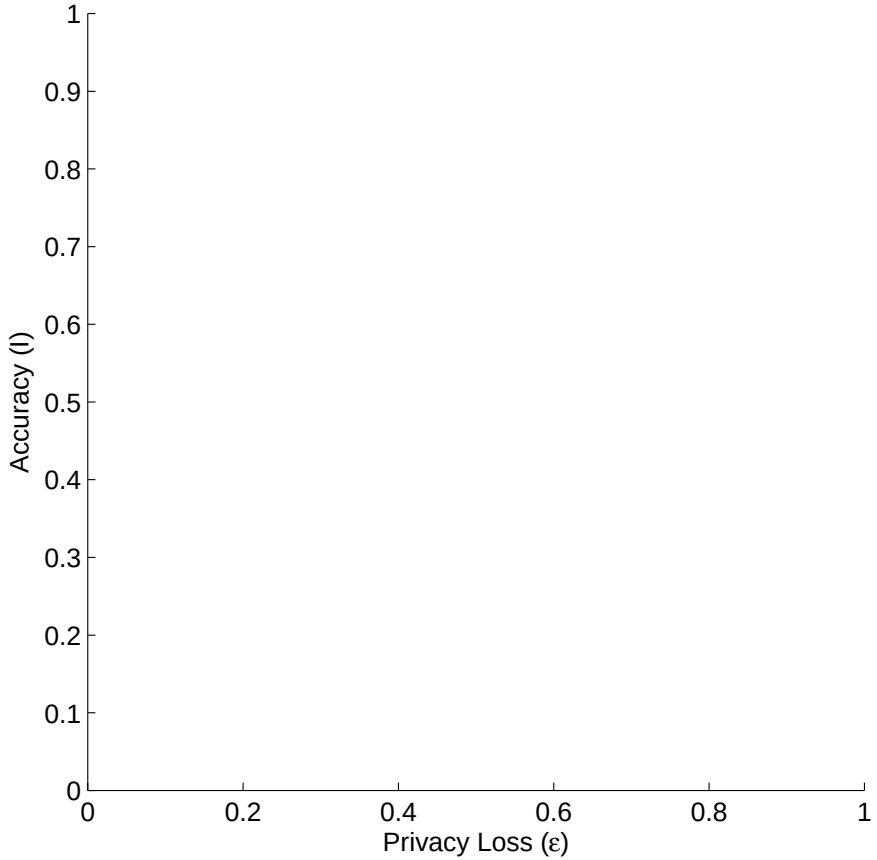
PWM is (ε, δ)-differentially private.

Differential Privacy and Inferential Disclosure

- Differential privacy is a property of a mechanism
- ε measures "worst-case" privacy loss
- Intuition: mechanism should behave similarly on neighboring databases
- Intuition: There can be no information from public-use statistics without some privacy loss

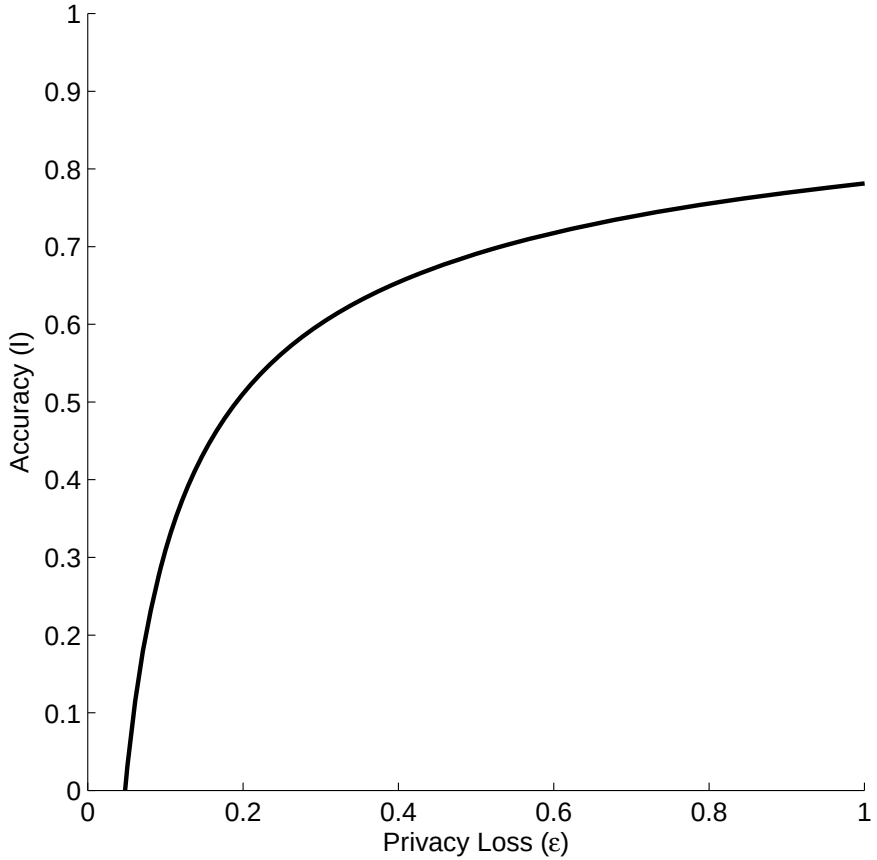
Production Possibilities under PMW

$$MRT(\varepsilon, I) \equiv \frac{dI}{d\varepsilon} = -\frac{\partial G/\partial \varepsilon}{\partial G/\partial I} = \frac{bK(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^{b+1}} \tag{1}$$



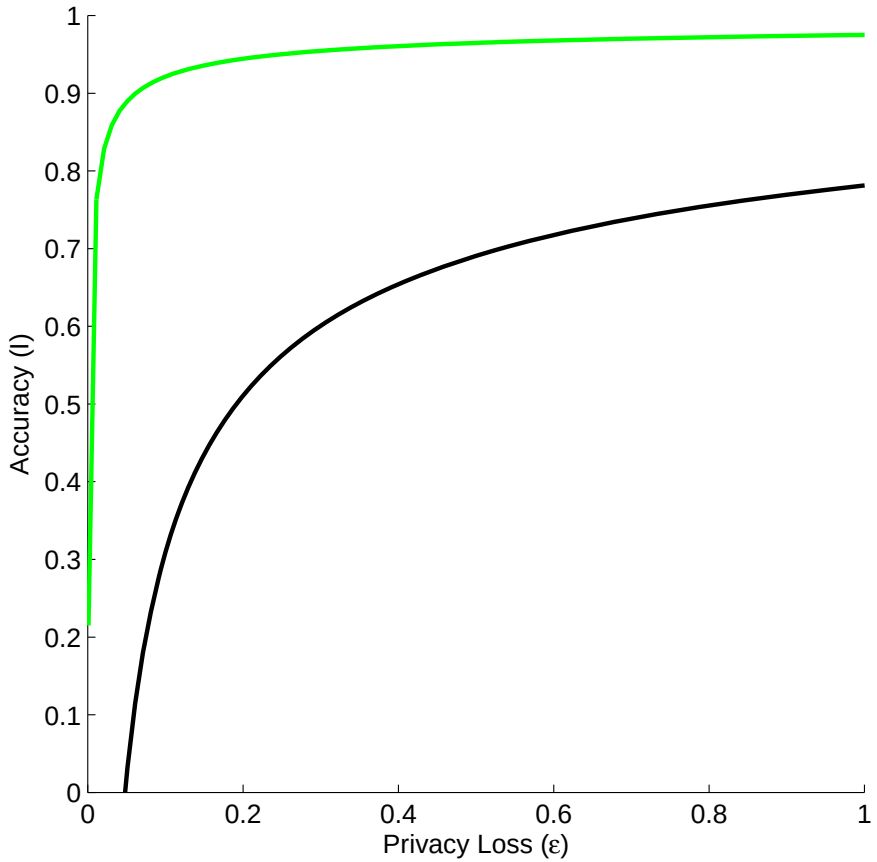
Production Possibilities under PMW

$$MRT(\varepsilon, I) \equiv \frac{dI}{d\varepsilon} = -\frac{\partial G/\partial \varepsilon}{\partial G/\partial I} = \frac{bK(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^{b+1}} \tag{1}$$



Production Possibilities under PMW

$$MRT(\varepsilon, I) \equiv \frac{dI}{d\varepsilon} = -\frac{\partial G/\partial \varepsilon}{\partial G/\partial I} = \frac{bK(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^{b+1}} \tag{1}$$



Preferences

$$
\begin{aligned}
v\left(y_i, \varepsilon, I, \tilde{y^i}, p\right) &= -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i \\
&\quad -\gamma_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) \varepsilon \\
&\quad +\eta_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) I
\end{aligned} \tag{2}
$$

- $\gamma_i > 0$ measures the preference for privacy
- $\eta_i > 0$ measures the preference for accuracy

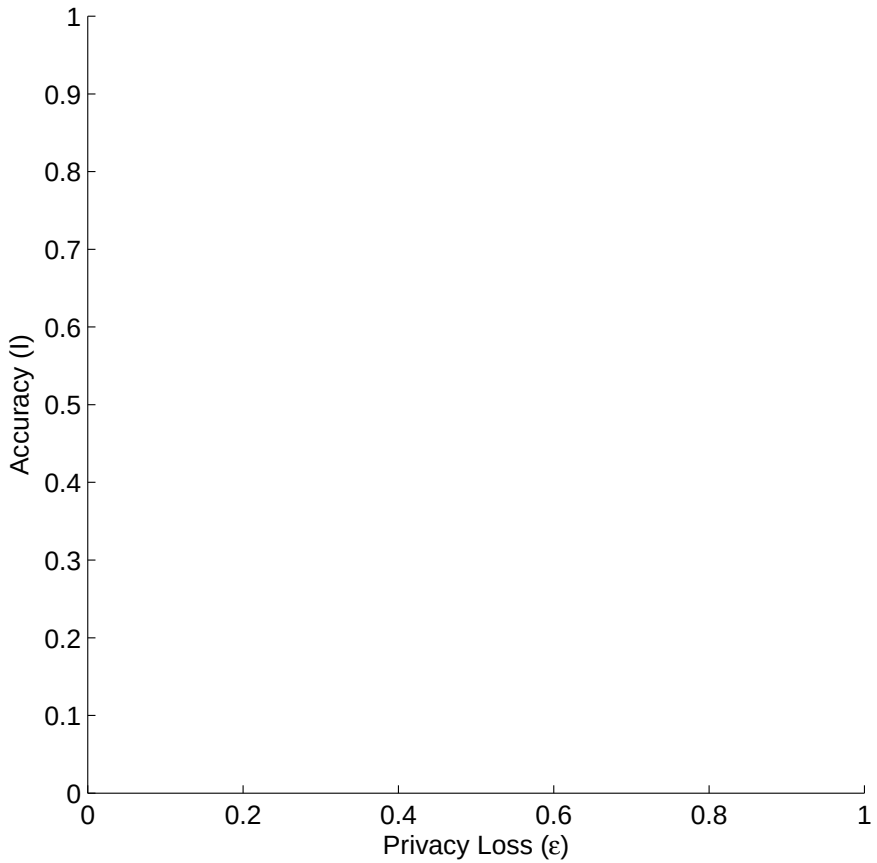
Accuracy and Privacy are Public Goods

In this setting

- Accuracy, I, is a public good
  - No privileged access to the data
- Differential privacy, ε, is a public bad
  - equal protection of all citizens

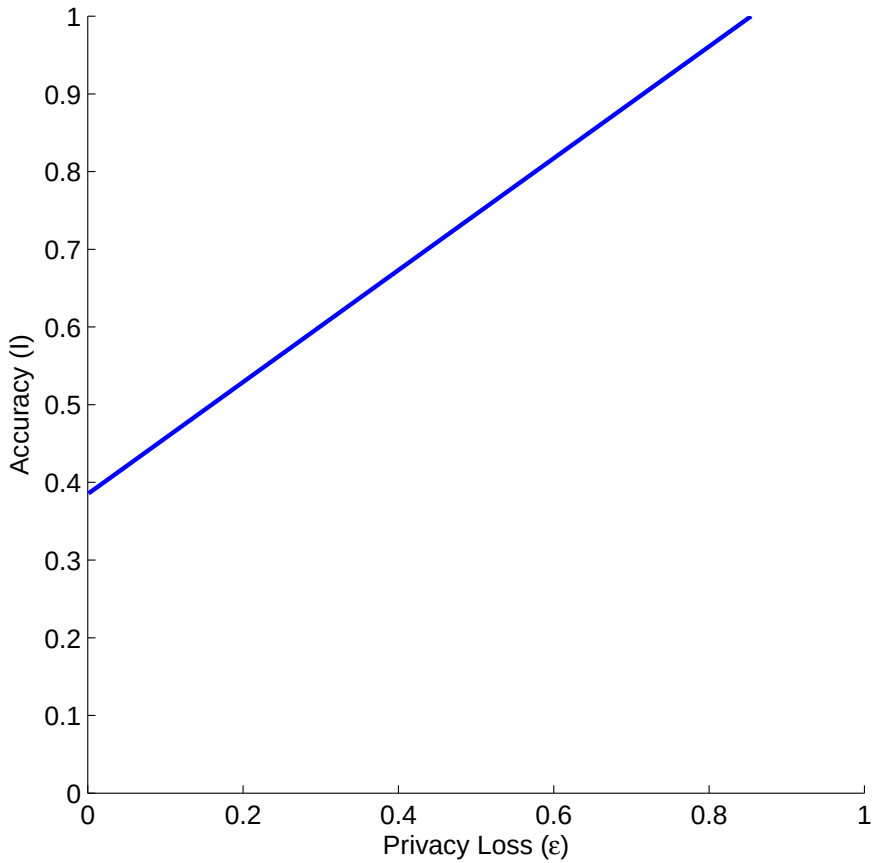
Social Welfare Function

$$SWF\left(\varepsilon, I, v, y, p\right) = \sum_{i=1}^{N} v_i\left(y_i, \varepsilon, I, \tilde{y^i}, p\right) \tag{3}$$



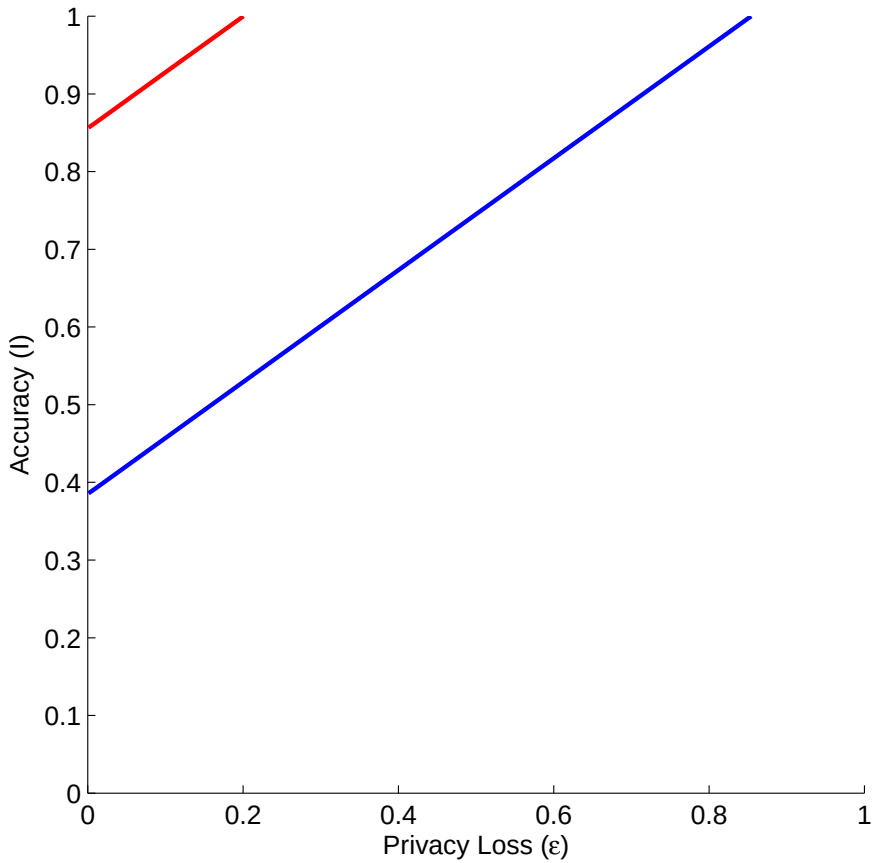
Social Welfare Function

$$SWF\left(\varepsilon, I, v, y, p\right) = \sum_{i=1}^{N} v_i\left(y_i, \varepsilon, I, \tilde{y^i}, p\right) \tag{3}$$



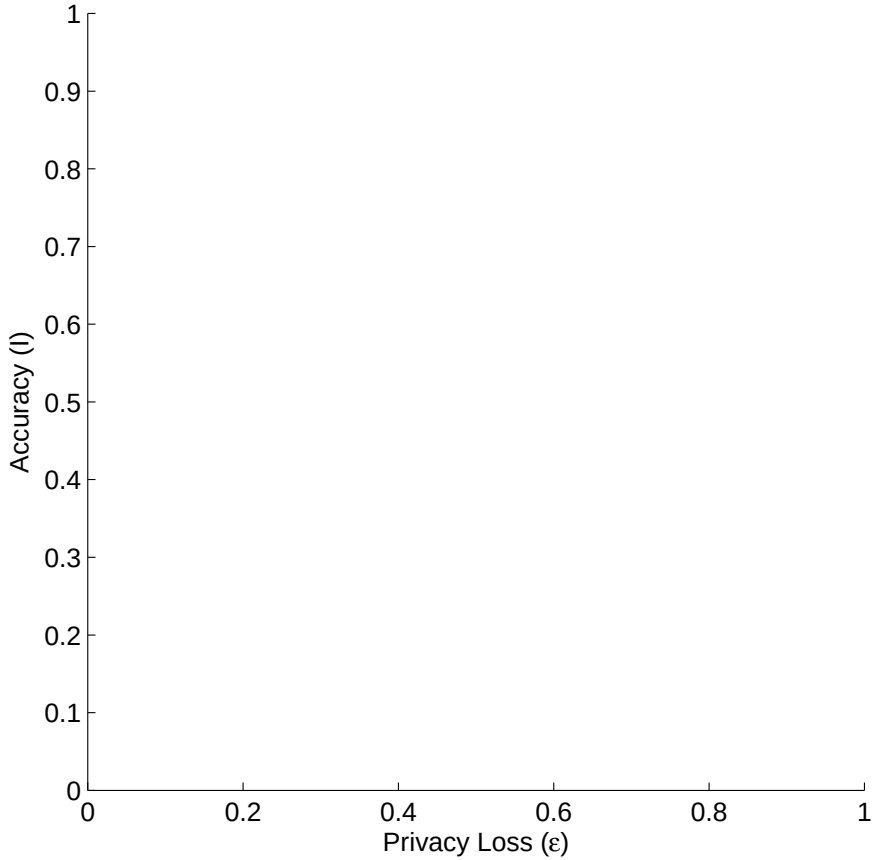
Social Welfare Function

$$SWF\left(\varepsilon, I, v, y, p\right) = \sum_{i=1}^{N} v_i \left(y_i, \varepsilon, I, \tilde{y^i}, p\right) \tag{3}$$



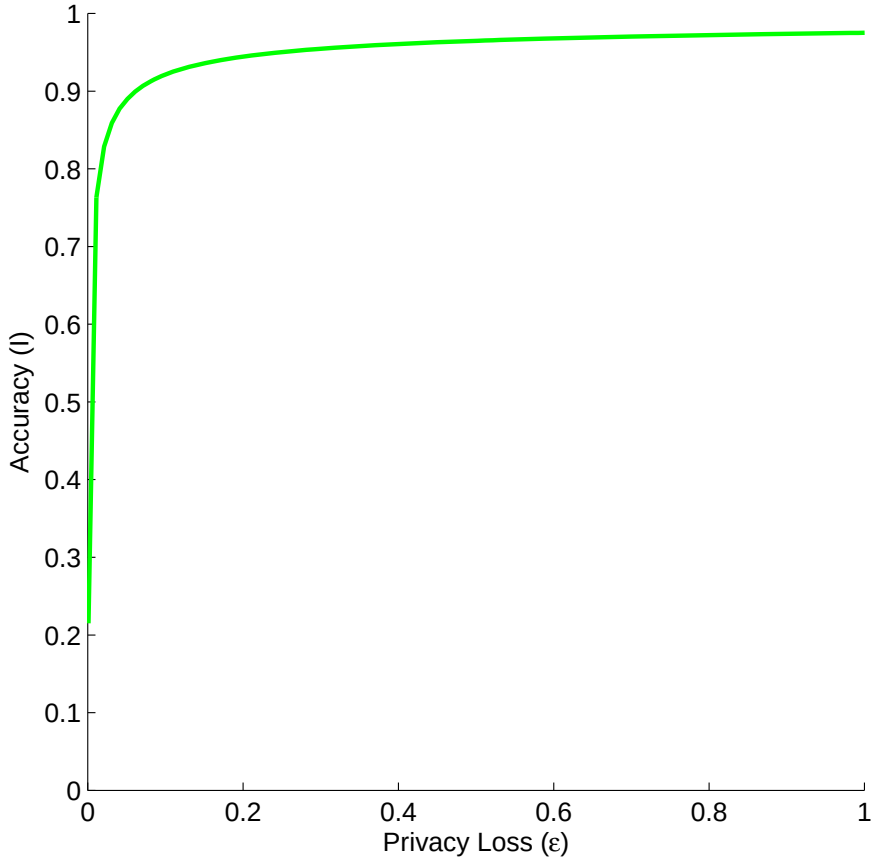
Social Welfare Maximization

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



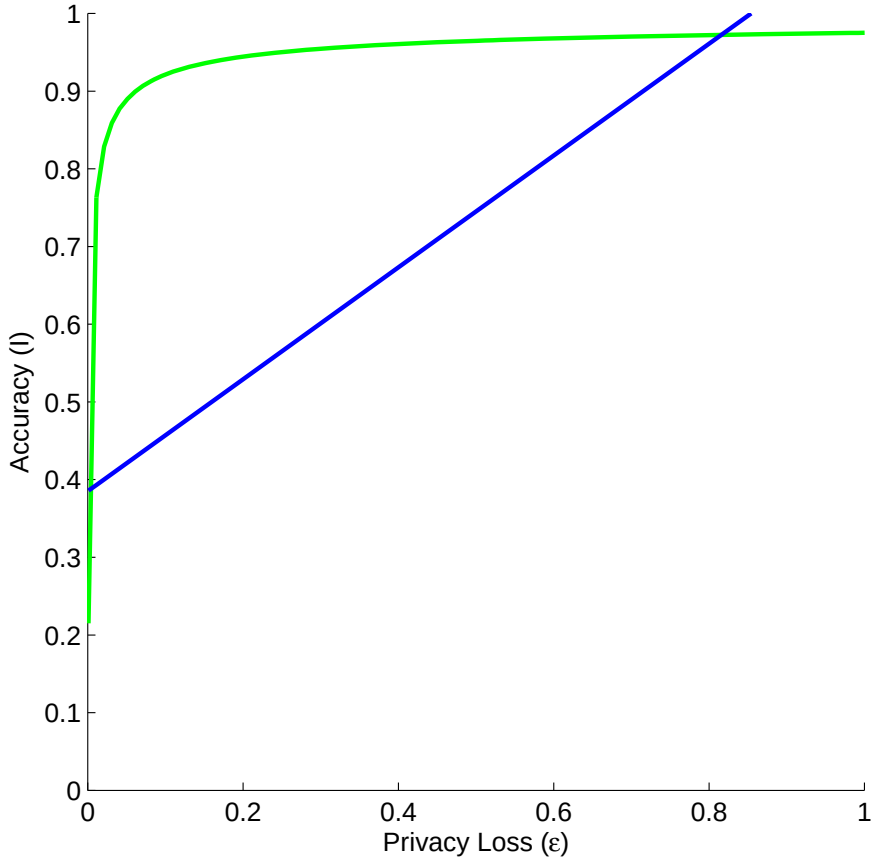
Social Welfare Maximization

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



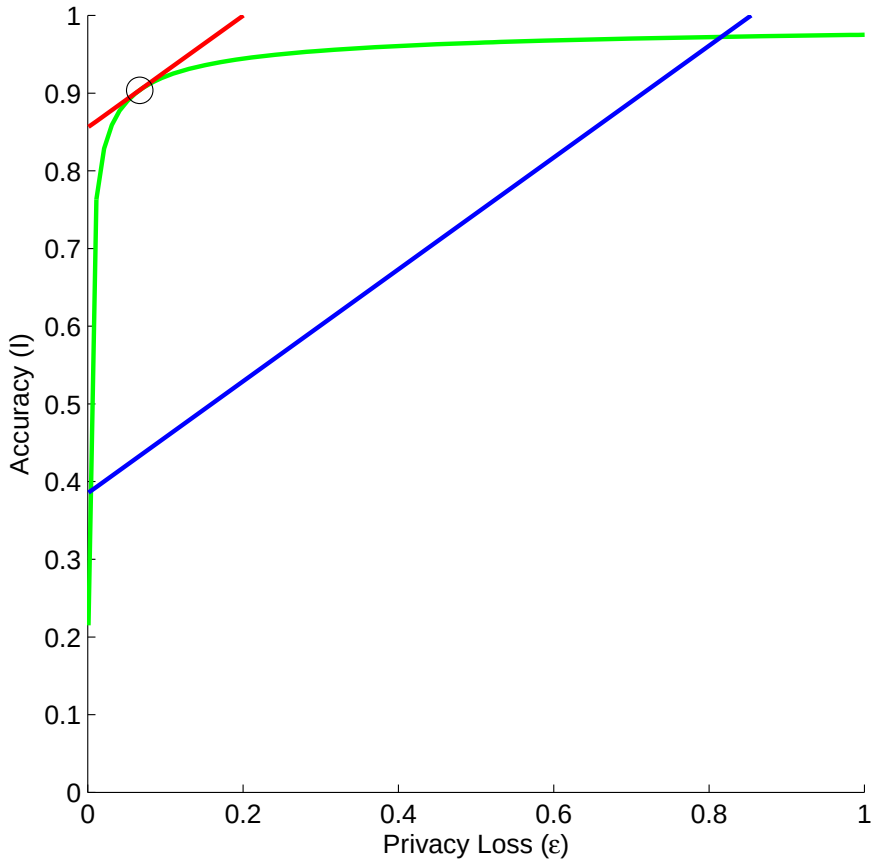
Social Welfare Maximization

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



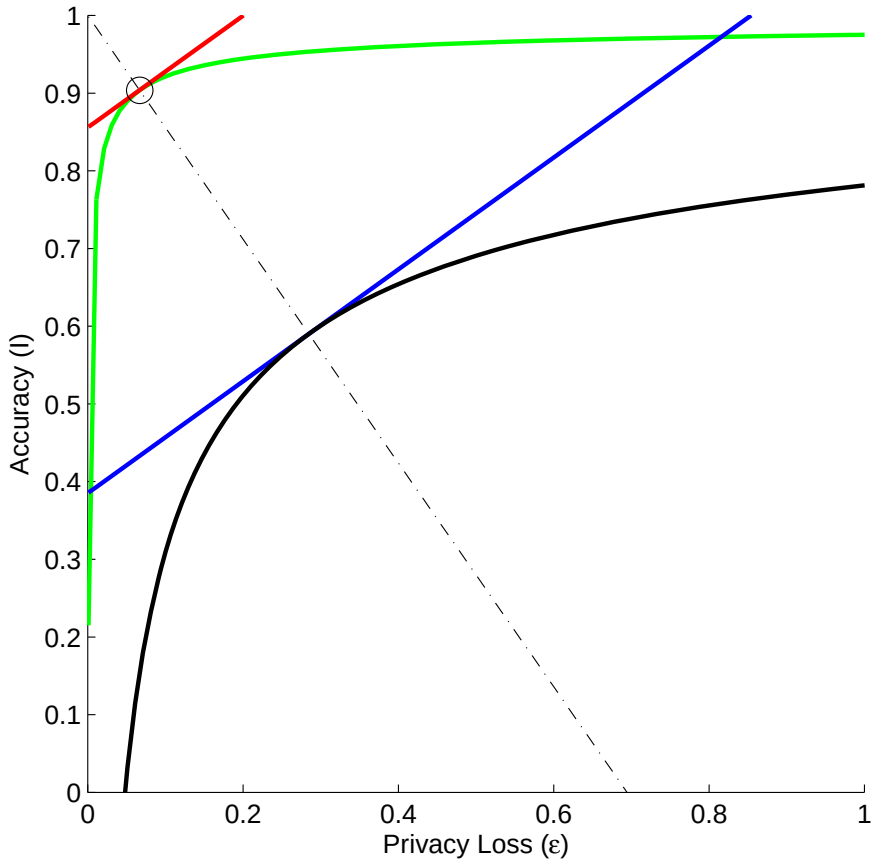
Social Welfare Maximization

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



Social Welfare Maximization

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



Solution

$$MRT = WTP \tag{4}$$

$$\begin{aligned} \frac{\frac{\partial G\left(\varepsilon^0, I^0\right)}{\partial \varepsilon}}{\frac{\partial G\left(\varepsilon^0, I^0\right)}{\partial I}} &= \frac{\frac{\partial}{\partial \varepsilon} \sum_{i=1}^{N} v_i\left(y_i, \varepsilon^0, I^0, \tilde{y^i}, p\right)}{\frac{\partial}{\partial I} \sum_{i=1}^{N} v_i\left(y_i, \varepsilon^0, I^0, \tilde{y^i}, p\right)} \\ \frac{bK(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{\left(\varepsilon^0\right)^{b+1}} &= \frac{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]}{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]} \end{aligned} \tag{5}$$

- Optimal choice caters to the average consumer

Application: Publication of Income Statistics

- The population in the database is the 2010 population of the United States ages 18 to 64: $N = 194{,}000{,}000$
- The size of the sample space is the number of bins into which the cardinal discrete variable is tabulated: $|\chi| = 1{,}000$
- The size of the query set is the number of permissible queries of the allowable form: $|\mathcal{Q}| = |\chi| - 1 = 999$
- The probability that a query cannot be answered accurately is $\beta = 0.01$
- The differential privacy failure bound is $\delta = 0.9/N$

Measuring Preferences 1

Three variables from the 2006 GSS:

- Family income, reported in quartiles.
- `DATABANK`, proxy for latent preference for privacy γ_i
  "The federal government has a lot of different pieces of information about people which computers can bring together very quickly. Is this a very serious threat to individual privacy, a fairly serious threat, not a serious threat, or not a threat at all to individual privacy?"
- `SCIIMP3`: proxy for latent preference for accuracy η_i.
  "How important [is] the following in making something scientific? The conclusions are based on solid evidence."

Polychoric correlations:

- $\text{Corr}\left[\gamma_i, \ln y_i\right] = -0.144\,(\pm 0.031)$
- $\text{Corr}\left[\eta_i, \ln y_i\right] = 0.189\,(\pm 0.037)$

Solution using GSS Preferences

$$WTP\left(\varepsilon^0, I^0\right) = \frac{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]}{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]} = 0.720. \tag{6}$$

- Optimal choices:
  - $I^* = 0.9$
  - $\varepsilon^* = 0.07$

- Welfare loss: a 3 percent decrease in accuracy requires an 0.8 percent increase in income

Measuring Preferences 2

We use five variables from the ongoing Center for Survey Methodology FSS Trust Survey:

- Family income, reported in 5 categories.
- `FS11`, proxy for latent preference for privacy γ_i
  "People can trust federal statistical agencies to keep information about them confidential. [5 categories]"
- `FS14a`, proxy for latent preference for privacy γ_i
  "Would you say that federal statistical agencies often invade people's privacy, or generally respect people's privacy? [2 categories]"
- `FS6`: proxy for latent preference for accuracy η_i.
  "Personally, how much trust do you have in the federal statistics in the United States? Would you say that you tend to trust federal statistics or tend not to trust them? [2 categories]"

Measuring Preferences 2 II

- `FS7`: proxy for latent preference for accuracy η_i.
  "Policy makers need federal statistics to make good decisions about things like federal funding. [5 categories]"
- `FS8`: proxy for latent preference for accuracy η_i.
  "Statistics provided by federal agencies are generally accurate. [5 categories]"

Application: Income Statistics, CSM data

Polychoric correlations:

- $\mathrm{Corr}\left[\gamma_i, \ln y_i\right] = 0.020\,(\pm 0.003)$ from FS11
- $\mathrm{Corr}\left[\gamma_i, \ln y_i\right] = 0.113\,(\pm 0.007)$ from FS14a
- $\mathrm{Corr}\left[\eta_i, \ln y_i\right] = 0.119\,(\pm 0.004)$ FS6
- $\mathrm{Corr}\left[\eta_i, \ln y_i\right] = 0.094\,(\pm 0.003)$ FS7
- $\mathrm{Corr}\left[\eta_i, \ln y_i\right] = 0.076\,(\pm 0.004)$ FS8

Application: Income Statistics, data from the CSM FSS Survey

- WTP = 0.914 ($\pm$0.005) Privacy: FS11, Accuracy FS6
- WTP = 0.998 ($\pm$0.007) Privacy: FS14a, Accuracy FS6
- WTP = 0.932 ($\pm$0.004) Privacy: FS11, Accuracy FS7
- WTP = 1.017 ($\pm$0.007) Privacy: FS14a, Accuracy FS7
- WTP = 0.947 ($\pm$0.005) Privacy: FS11, Accuracy FS8
- WTP = 1.034 ($\pm$0.008) Privacy: FS14a, Accuracy FS8

Other Results: Private Provision is Suboptimal

Based on Ghosh and Roth (2011)

1. ε-DP can be priced through a procurement auction
2. A VCG auction yields minimum-cost method for answering a query with
   - $(\varepsilon, 0)$-differential privacy
   - (α, β)-accuracy
3. Our extension:
   - Model the supply of accuracy by a private firm
   - Sells $\hat{s}$ with data quality I at price per unit of quality, p
   - **Key Result**

Theorem

$I^{VCG} \leq I^{L} \leq I^{0}$, where I^{0} is the Pareto optimal level of I , I^{L} is the privately-provided level when using the Lindahl mechanism to procure data-use rights and I^{VCG} is the privately-provided level when using the VCG procurement mechanism

Future Work

- Better data
  - Census Bureau survey on privacy and accuracy attitudes
  - Experimental measures of preferences
- Model improvements
  - Technology: implement PMW and other mechanisms in real data
  - Measure PPF in real-world use cases
  - Demand for privacy
  - Demand for accuracy (public capital)