A Theory of Pricing Private Data

Dan Suciu – U. of Washington

Joint work with: Chao Li, Daniel Yang Li, Gerome Miklau

Motivation

- Private data has value
  - A unique user: $4 at FB, $24 at Google [JPMorgan]
- Today's common practice:
  - Companies profit from private data without compensating users
- New trend: allow users to profit financially
  - Industry: personal data locker https://www.personal.com/ , http://lockerproject.org/
  - Academia: mechanisms for selling private data [Ghosh11,Gkatzelis12,Aperjis11,Roth12,Riederer12]

Overview

This talk: framework for pricing queries on private data

- Data owners: sell their private data
- Buyer: buys a query (many buyers, many queries!)
- Trusted market maker: facilitates transactions

What I will address:

- Consistent prices for arbitrary queries
- Fair compensation of data owners for privacy loss

What I will not address:

- Designing truthful, efficient mechanisms
- Prices/payments: at the discretion of market maker

Challenges

Perturbation: is a cost savings mechanism for buyer
Price: computed for each (query, perturbation) pair.

Two extremes:

- No perturbation
  - Query returns raw data
  - Data owner compensated the full price of data; e.g. $10
  - Buyer pays a high price
- High perturbation
  - Query is ε-Differentially Private, for small ε
  - Data owner compensated a tiny price, e.g. $0.001
  - Buyer pays modest price

Related Work

- Query-based data pricing, Koutris, Upadhyaya, Balazinska, Howe, Suciu, 2012
- Pricing Aggregate Queries in a Data Marketplace, Li and Miklau, 2012
- Selling privacy at auction, Ghosh, A., Roth, A. 2011
- Pricing Private Data, Gkatzelis, Aperjis, Huberman, 2012
- A Market for Unbiased Private Data, Aperjis, Huberman 2011
- Buying Private Data at Auction (…), Roth 2012
- For sale : Your Data By : You, Riederer, Erramilli, Chaintreau, Krishnamurthy, Rodriguez, 2012

Outline

- $\boxed{\text{Problem Statement}}$
- The Buyer's price: π
- Balanced Pricing Framework
- Conclusions

Main Concepts

- Database $\mathbf{x} = (x_1, \ldots, x_n)$
  - x_i = *value,* owned by some *owner*
- Buyer's request: $\mathbf{Q} = (\mathbf{q}, v)$
  - $\mathbf{q} = (q_1, \ldots, q_n)$ = query; $\mathbf{q}(\mathbf{x}) = \Sigma_i q_i x_i$
  - v = variance
- Randomized answer: $\mathcal{K}(\mathbf{x})$
  - $E[\mathcal{K}(\mathbf{x})] = \mathbf{q}(\mathbf{x})$, $Var[\mathcal{K}(\mathbf{x})] \leq v$

Buyer pays $\pi(\mathbf{Q})$

- Privacy loss:
  - $\varepsilon_i(\mathcal{K})$ [Ghosh'11]
  - $W(\varepsilon_i)$ = its value to the owner

Owner receives $\mu_i(\mathbf{Q})$

Example (1/3)

Data: 1000 data owners rate two candidates A, B between 0..5:

- Owner 1: x_1, x_2
- Owner 2: x_3, x_4
- …
- Owner 1000: x_{1999}, x_{2000}

Price: \$10 for each raw item x_i

- Buyer:
  - Compute rating for candidate A: $x_1+x_3+\ldots+x_{1999}$
  - **q** = (1,0,1,0,…), v=0 (raw data)
- μ-Payments: \$10/item
- Buyer's Price π: \$10,000

1. Raw data is too expensive!

Example (2/3)

Data: 1000 data owners rate two candidates A, B between 0..5:

- Owner 1: x_1, x_2
- Owner 2: x_3, x_4
- …
- Owner 1000: x_{1999}, x_{2000}

Price: \$10 for each raw item x_i

- Buyer:
  - Can tolerate error ±300
  - **q** = (1,0,1,0,…), ~~v=0~~ v = 2500* ($v=\sigma^2$ = variance)
- μ-Payments: ~~\$10/item~~ \$0.001/item (query is 0.1-DP**)
- Buyer's Price π: ~~\$10,000~~ \$1

2. Perturbed data is cheaper.

*Probability(error < 6σ) > $1/6^2$ = 97%

** ε = Sensitivity(**q**)/σ = 5/σ = 0.1

Example (3/3)

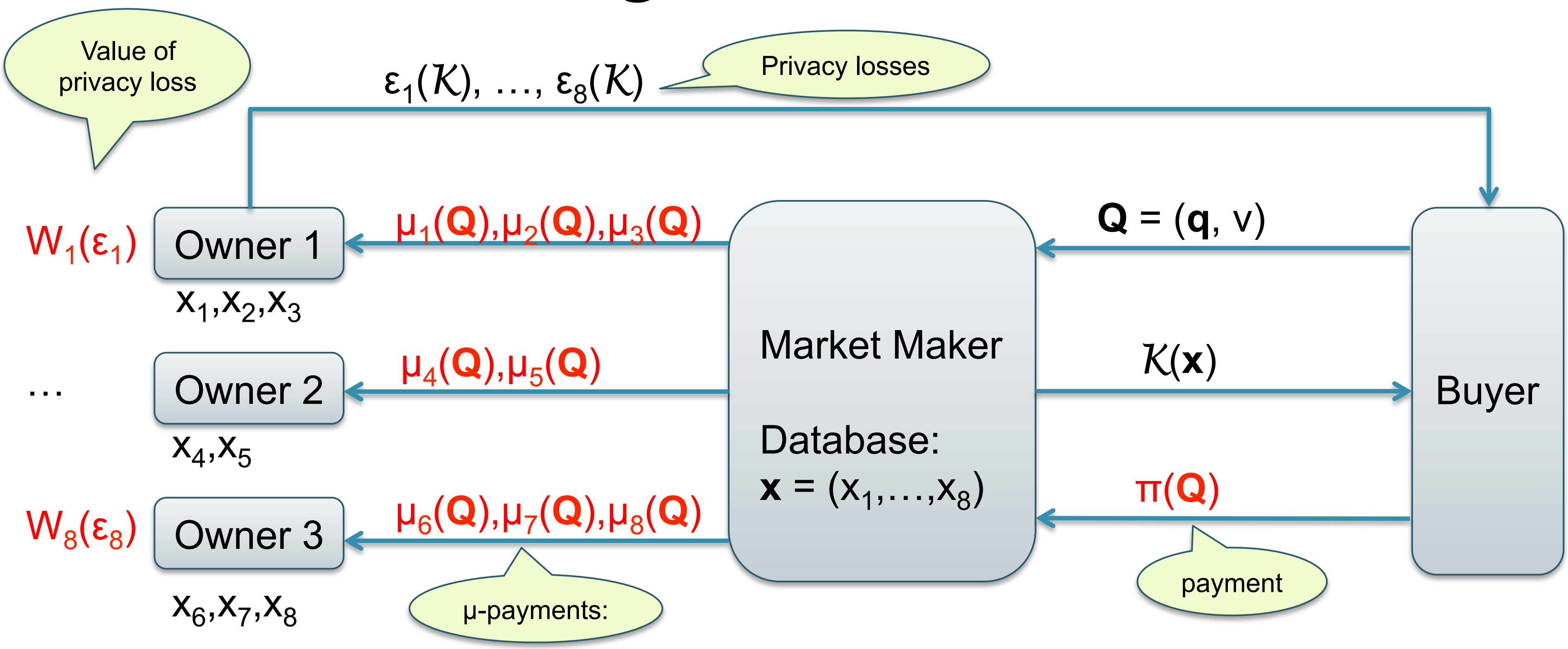
Data: 1000 data owners rate two candidates A, B between 0..5:
- Owner 1: x_1, x_2
- Owner 2: x_3, x_4
- …
- Owner 1000: x_{1999}, x_{2000}

Price: \$10 for each raw item x_i

- Another buyer:
  - **q** = (1,0,1,0,…), ~~variance = 0, variance = 2500~~ variance = 500
- μ-Payments: ~~\$10/item,\$0.001/item~~ \$0.1/item? \$1/item?
- Buyer's Price π: ~~\$10000, \$1~~ \$100? \$1000?
- Buyer will refuse to pay more than \$5!
  - Instead purchases 5 times variance=2500, for \$5, takes avg.

3. Multiple queries: must be consistent, compensate owners for privacy loss.

Pricing Framework

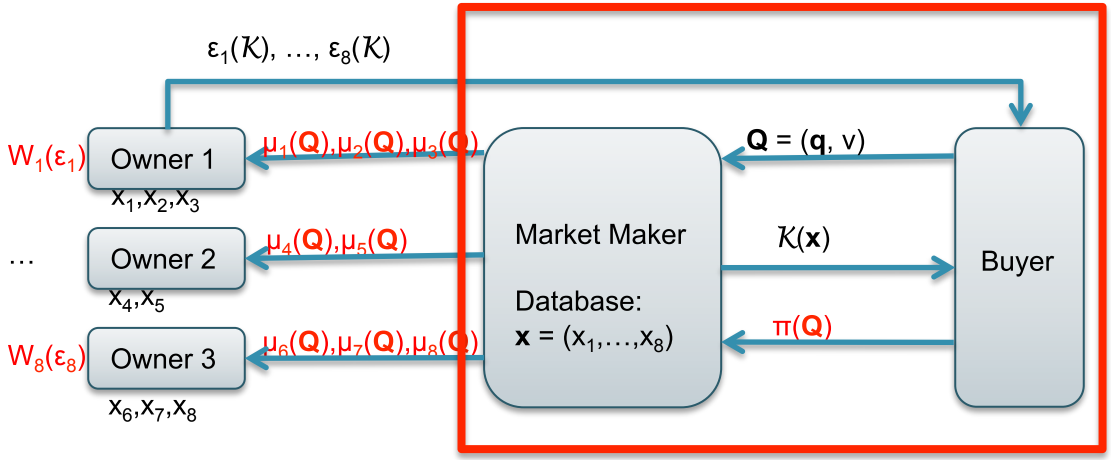


Market maker needs to balance the pricing framework

- Satisfy the buyer: use $\mathcal{K}$ to answer **Q**, charge him $\pi(\mathbf{Q})$
- Satisfy the owner: pay her $\mu_i(\mathbf{Q}) \geq W_i(\varepsilon_i)$
- Recover cost: $\mu_1 + \ldots + \mu_n \leq \pi$

Outline

- Problem Statement
- The Buyer's price: π
- Balanced Pricing Framework
- Conclusions



Designing a Pricing Function

For any query/variance request $\mathbf{Q} = (\mathbf{q}, v)$

define a price: $\pi(\mathbf{Q}) \in [0, \infty]$

What can go wrong?

Arbitrage!

Def.

- $\mathbf{Q}=(\mathbf{q}, v)$ is *answerable* from $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$ $(=(\mathbf{q}_1 v_1), \ldots, (\mathbf{q}_k v_k))$ if there exists a function f s.t. whenever $\mathcal{K}_1, \ldots, \mathcal{K}_k$ answer $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$, $f(\mathcal{K}_1, \ldots, \mathcal{K}_k)$ answers $\mathbf{Q}$
- $\mathbf{Q}$ is *linearly answerable* from $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$ if f is a linear function; notation: $\mathbf{Q}_1, \ldots, \mathbf{Q}_k \rightarrow \mathbf{Q}$

Examples: $(\mathbf{q}_1,v_1), (\mathbf{q}_2,v_2), (\mathbf{q}_3,v_3) \rightarrow (\mathbf{q}_1+\mathbf{q}_2+\mathbf{q}_3, v_1+v_2+v_3)$

$(\mathbf{q}, v) \rightarrow (c\,\mathbf{q}, c^2 v)$

$(\mathbf{q},v), (\mathbf{q},v), (\mathbf{q},v), (\mathbf{q},v), (\mathbf{q},v) \rightarrow (\mathbf{q},v/5)$

Def. *Arbitrage* happens when $\mathbf{Q}_1, \ldots, \mathbf{Q}_k \rightarrow \mathbf{Q}$ and $\pi(\mathbf{Q}_1) + \ldots + \pi(\mathbf{Q}_k) < \pi(\mathbf{Q})$

Example: If $5\times\pi(\mathbf{q},v) < (\mathbf{q},v/5)$, then we have aribtrage

Arbitrage-Free Pricing

Def. The pricing function π is *Arbitrage–Free* if:
$\mathbf{Q}_1, \ldots, \mathbf{Q}_k \rightarrow \mathbf{Q}$ implies $\pi(\mathbf{Q}_1) + \ldots + \pi(\mathbf{Q}_k) \geq \pi(\mathbf{Q})$

Do AF-pricing functions exists?

Remark: AF generalizes the following known property of ε-DP:

If $\mathbf{Q}_1$ is ε-DP, and $\mathbf{Q} = f(\mathbf{Q}_1)$, then $\mathbf{Q}$ is also ε-DP

Indeed: if $\pi(\mathbf{Q}_1) \leq \$0.001$ then $\pi(\mathbf{Q}) \leq \$0.001$

Designing Arbitrage-Free Pricing Functions

$\pi(\mathbf{q}, v) = (q_1^2 + q_2^2 + \ldots + q_n^2) / v$ is AF

Price of raw data $\pi(\mathbf{q}, 0) = \infty$

More generally:
$\pi(\mathbf{q}, v) = \| \mathbf{q} \|^2 / v$ is AF, where $\| \mathbf{q} \|$ is any *semi-norm*

$\pi(\mathbf{q}, v) = 20{,}000 / 3.14 \times \arctan[(q_1^2 + q_2^2 + \ldots + q_n^2) / v]$

Price of raw data $\pi(\mathbf{q}, 0) = 10{,}000$

More generally:
If f is sub-additive, non-decreasing and $\pi_1, \ldots, \pi_k$ are AF
then $\pi = f(\pi_1, \ldots, \pi_k)$ is AF

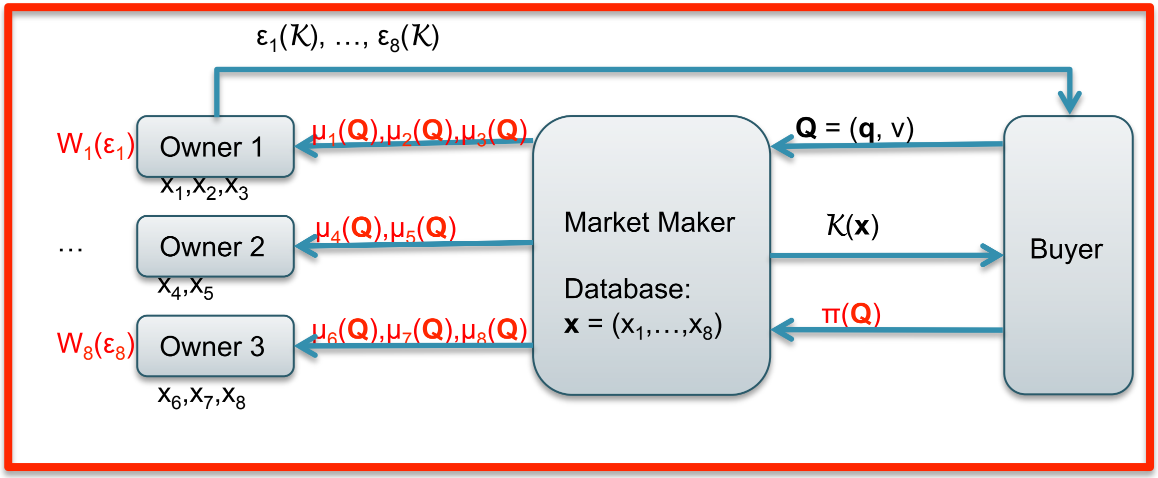
Discussion

- *Query answerability* is well studied for relational queries (no noise!) [Nash'2010]
  - Checking answerability: NP … undecidable

- New for linear queries with noise:
  - Checking linear answerability is in PTIME
  - Checking general answerability is open

Outline

- Problem Statement
- The Buyer's price: π
- Balanced Pricing Framework
- Conclusions



The Perspective of the Data Owner

- Micropayment to owner i:
  $\mu_i(\mathbf{Q})$ = what the market maker pays her

- Must compensate for her privacy loss: [Ghosh'11]

$$\varepsilon_i(\mathcal{K}) = sup_{S,\mathbf{x}} \left| \log \frac{\Pr\left(\mathcal{K}(\mathbf{x}) \in S\right)}{\Pr\left(\mathcal{K}(\mathbf{x}^{(i)}) \in S\right)} \right|$$

  $W_i(\varepsilon_i)$ = the owner's value for the privacy loss

  $W_i(\infty)$ = price for her raw data; e.g. = $10

Properties of μ_i

Assumptions: the pricing framework is defined by μ_i, W_i, plus:

- $\mathcal{K}$ = Laplacian answering mechanism:
  $\mathcal{K}(\mathbf{x}) = \mathbf{q}(\mathbf{x}) + \text{Lap}(\text{sqrt}(v/2))$
- $\pi = a(\mu_1 + \ldots + \mu_n) + b$, for some $a \geq 1$, $b \geq 0$

$\varepsilon_i(\mathcal{K})$ derived from sensitivity

market maker recovers the costs

Def. The pricing framework is *balanced* if is

(1) μ_i is arbitrage free,

(2) compensates owner: $\mu_i(\mathbf{Q}) \geq W_i(\varepsilon_i(\mathcal{K}))$

(3) is fair: $q_i = 0$ implies $\mu_i(\mathbf{q}, v) = 0$

Market maker must design a balanced pricing framework

Designing Balanced Pricing Frameworks

The pricing-frameworks below are balanced (assume $x_i \in [0,5]$)

$\mu_i(\mathbf{q}, v) = 5c_i\, |q_i| / \mathrm{sqrt}(v/2)$
$W_i(\varepsilon_i) = c_i\, \varepsilon_i$

Price of raw data:
$\mu_i(\mathbf{q}, 0) = W_i(\infty) = \infty$

c_i is any constant

$\mu_i(\mathbf{q}, v) = 20 / 3.14 \times \arctan(5c_i\, |q_i| / \mathrm{sqrt}(v/2))$
$W_i(\varepsilon_i) = 20 / 3.14 \times \arctan(c_i\, \varepsilon_i)$

Raw data:
$\mu_i(\mathbf{q}, 0) = W_i(\infty) = \$10$

More generally:
If $\mu_{i1}, \ldots, \mu_{ik}$ and $W_{i1}, \ldots, W_{ik}$ are balanced and f_i is non-decreasing, subadditive then $\mu_i = f(\mu_{i1}, \ldots, \mu_{ik})$, $W_i = f(W_{i1}, \ldots, W_{ik})$ are balanced

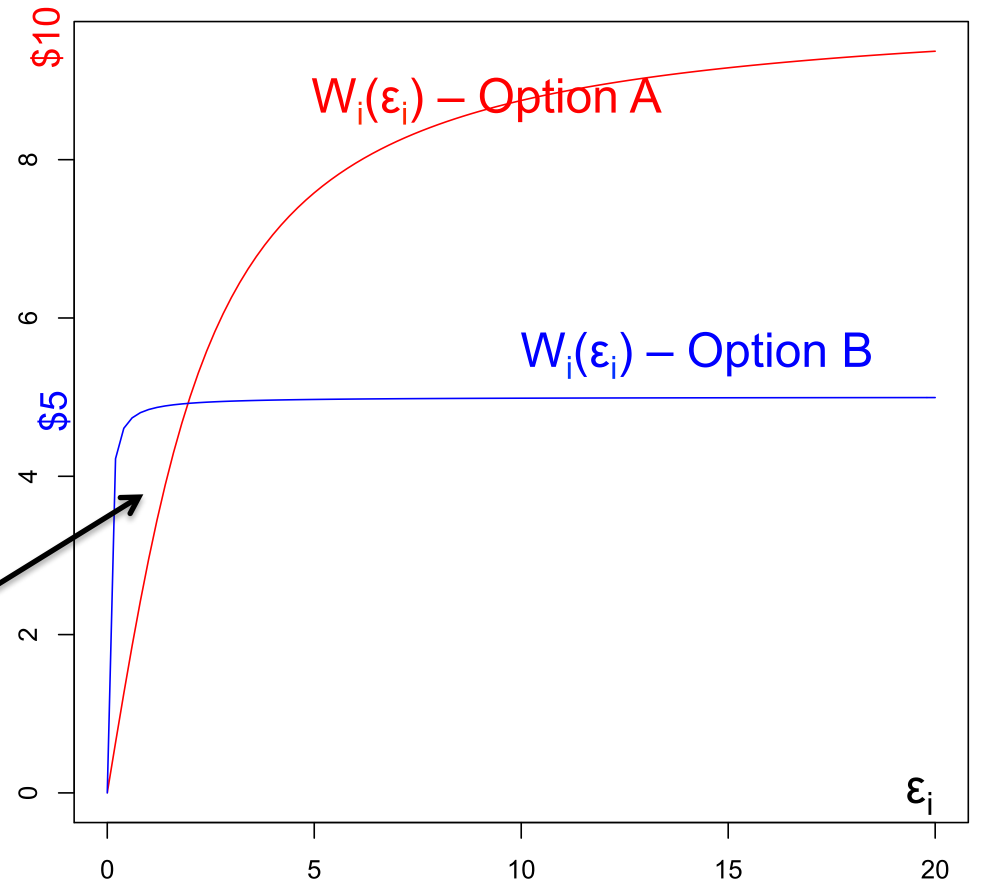
Finding Out the Owner's Valuation W_i

Mechanisms proposed [Ghosh'11,Gkatzelis'12,Riederer'12]
We use an idea from [Aperjis&Huberman'11]:

Market Maker
gives users 3 options

- Option A: risk neutral
- Option B: risk averse
- Option C: opt-out

"Typical" query has
small privacy loss



Outline

- Problem Statement
- The Buyer's price: π
- Balanced Pricing Framework
- Conclusions

Conclusions

- The Contract in differential-privacy:
  - *Privacy loss ε_i = bounded by a fixed, small ε*
  - **Privacy budget** (defined by ε) = limit on the number of queries
- The Contract in private data markets:
  - *Privacy loss ε_i = arbitrary; compensated by micro-payment μ_i*
  - **Cash-and-carry** = unlimited queries
- Special case 1: Answer contains raw data
- Special case 2: Answer is ε-DP
- Challenge: Designing a balanced pricing framework