Query-Based Data Pricing

Dan Suciu – U. of Washington

Joint with M. Balazinska, B. Howe, P. Koutris, Daniel Li, Chao Li, G. Miklau, P. Upadhyaya

Data Has Value

And it is increasingly being sold/bought on the Web

- Big data vendors
- Data Markets
- Private data

Pricing digital goods is challenging [Shapiro&Varian]

Pricing Data

Pricing data lies at the intersection of several areas:

- Data management
- Mechanism design
- Economics

This talk

1. Big Data Vendors

High value data

- Gartner report: $5k, even if you need only one chart
- Navteq Maps
- Factual
- A few others [Muschalle]:
  - Thomson Reuters, Mendeley Ltd., DataMarket Inc, Vico Research & Consulting GmbH, TEMIS S.A., Neofonie GmbH, Inovex GmbH

Expensive datasets, available only to major customers

2. Data Markets

- Azure DataMarkets – 100+ data sources
- Infochimps – 15,000 data sets
- Xignite – financial data
- Aggdata
- Gnip – social media data
- PatientsLikeMe

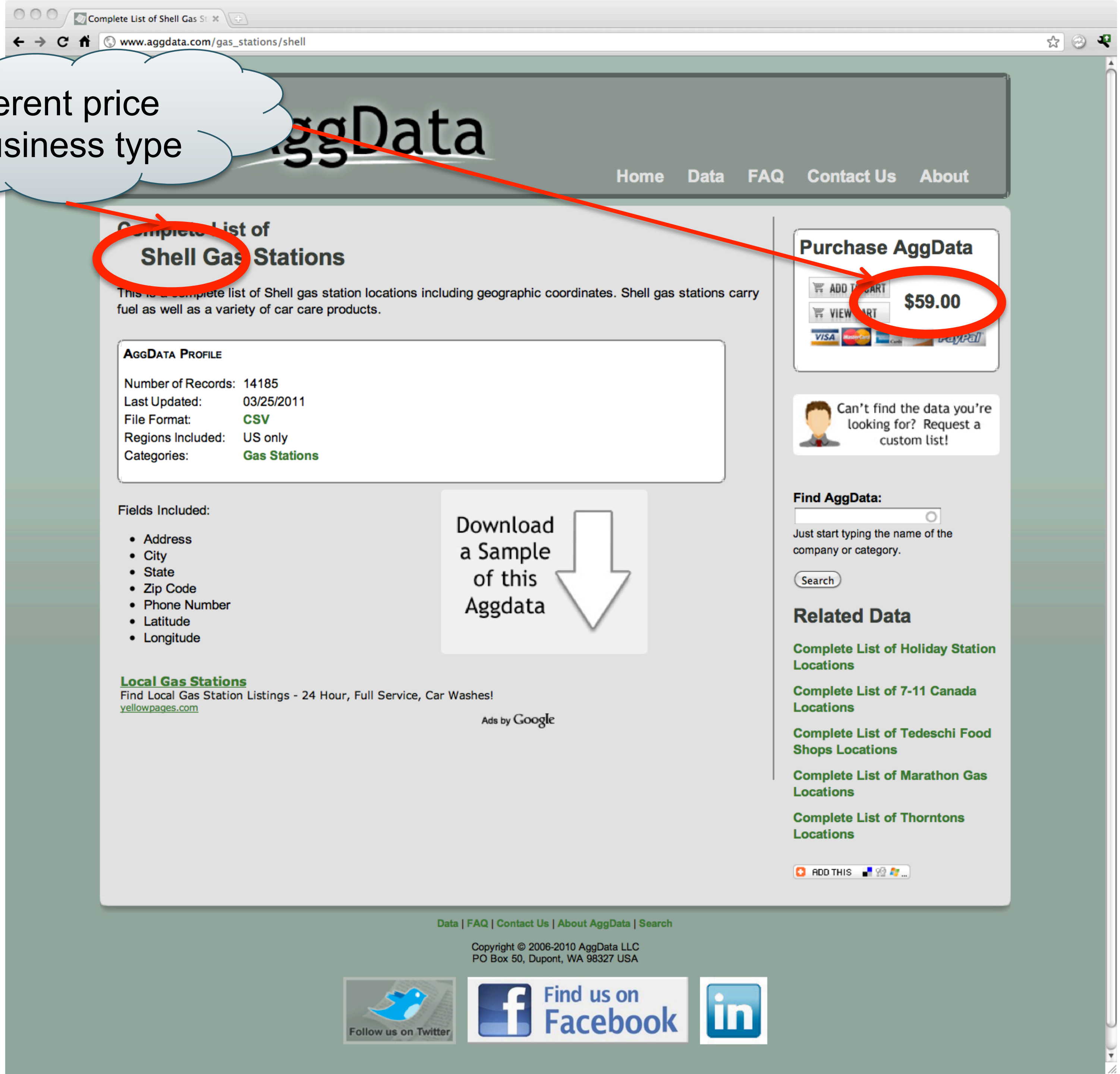
These datasets are available to the little guy.
The markets themselves are struggling,
because they are just facilitators; no innovation

3. Private Data

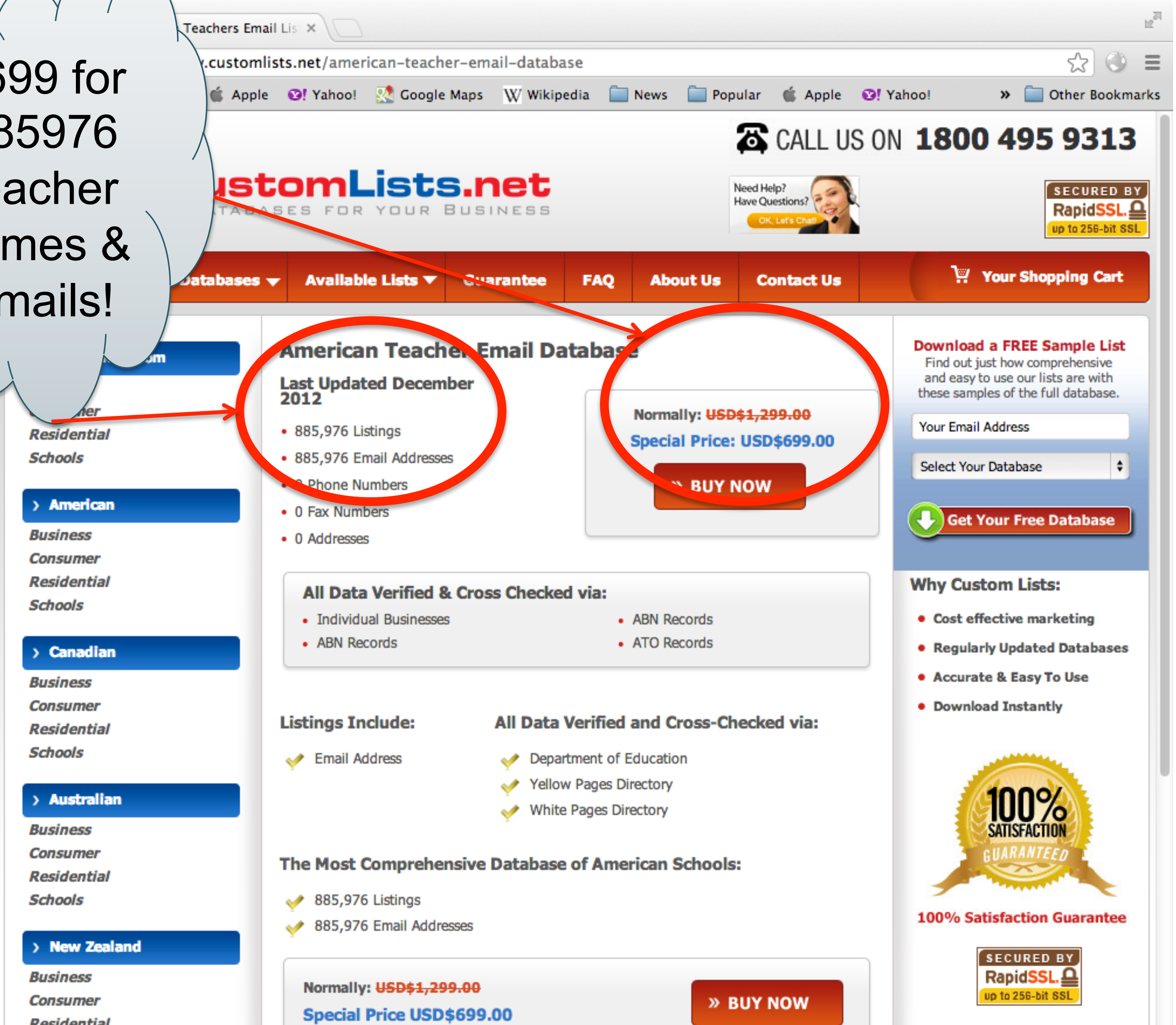
- Private data has value
  - A unique user: $4 at FB, $24 at Google [JPMorgan]
- Today's common practice:
  - Companies profit from private data without compensating users
- New trend: allow users to profit financially
  - Industry: personal data locker https://www.personal.com/ , http://lockerproject.org/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

Sample Data Markets

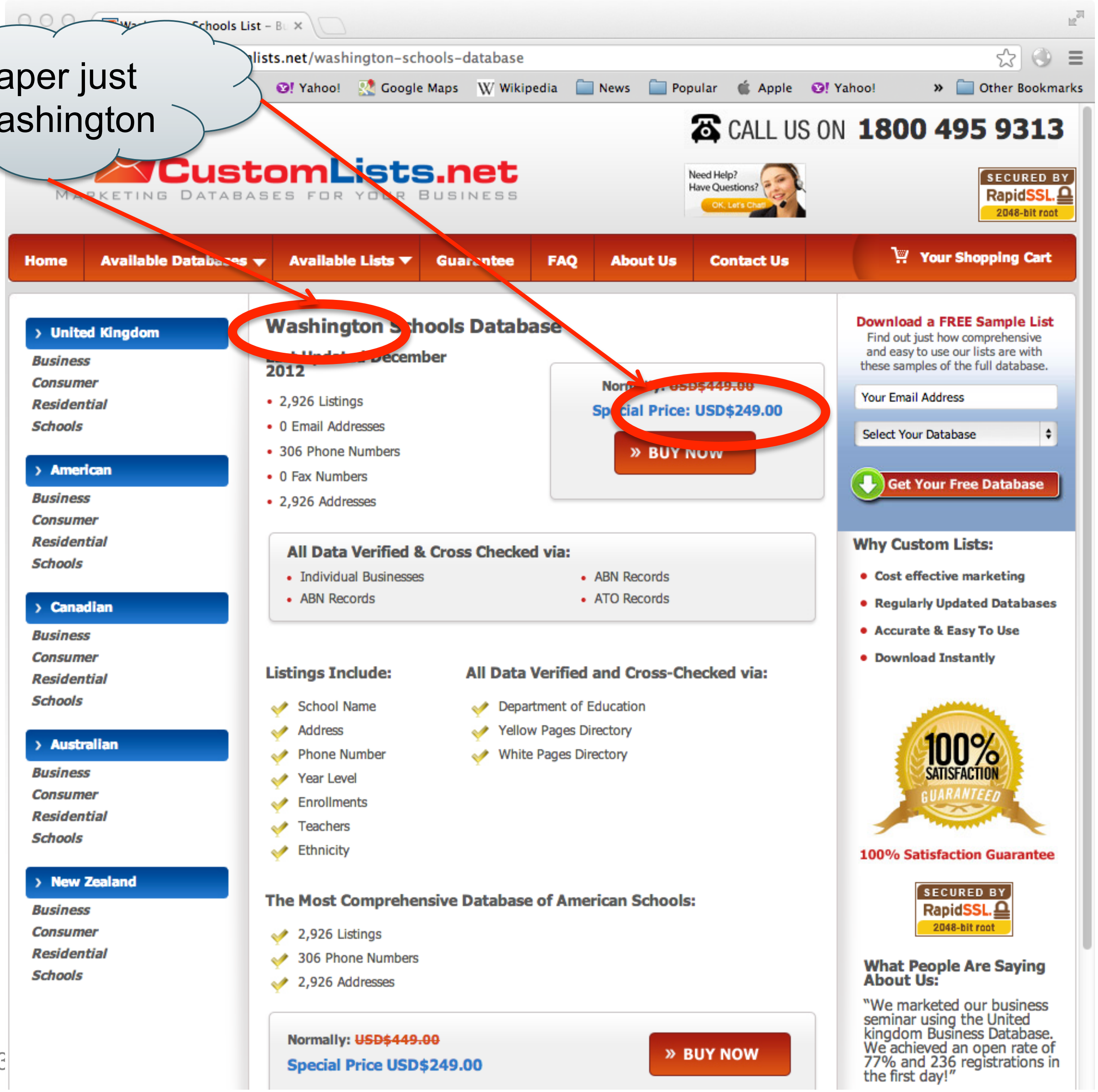
Different price by business type



$699 for 885976 teacher names & emails!



Cheaper just for Washington



A Criticism of Today's Pricing Schemes

- Small buyers want to purchase only a tiny amount of data: if they can't, they give up
- Large buyers have specific needs: price is often negotiated in a room-full-of-lawyers
- Sellers can't easily anticipate all possible queries that buyers might ask

Needed: more flexible pricing scheme, parameterized by queries

Outline

- Framework and examples
- Results so far
- Conclusions

Query-based Pricing

- Seller defines *price-points*:
  (V_1,p_1), (V_2, p_2), … Meaning: $\text{price}(V_i)=p_i$.

- Buyer may buy any *query* Q

- System will determine $\text{price}_D(Q)$ based on:
  - The price points
  - The current database instance D
  - The query Q

How should a "**good**" price function be?

Arbitrage Freeness

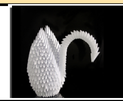
Arbitrage-free Axiom:
For all queries $Q_1, \ldots, Q_k$, Q,
if $Q_1, \ldots, Q_k$ *determine* Q, then:

$$\text{price}_D(Q) \leq \text{price}_D(Q_1) + \ldots + \text{price}_D(Q_k)$$

"$Q_1, \ldots, Q_k$ determine Q" means that
$Q(D)$ can be answered from $Q_1(D), \ldots, Q_k(D)$,
without accessing the database instance D

Example 1: Pricing Relational Data

S(Shape,Color,Picture)

Shape	Color	Picture
Swan	White	
Swan	Yellow	. . . . . .
Dragon	Yellow	
Car	Yellow	. . . . . .
Fish	White	. . . . . .

Price(σ_{Shape})=$2

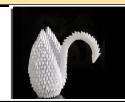
Price(σ_{Color})=$3

Price list	Price
$V_1 = \sigma_{Shape=\text{'Swan'}}(S)$	$2
$V_2 = \sigma_{Shape=\text{'Dragon'}}(S)$	$2
$V_3 = \sigma_{Shape=\text{'Car'}}(S)$	$2
$V_4 = \sigma_{Shape=\text{'Fish'}}(S)$	$2
$W_1 = \sigma_{Color=\text{'White'}}(S)$	$3
$W_2 = \sigma_{Color=\text{'Yellow'}}(S)$	$3
$W_3 = \sigma_{Color=\text{'Red'}}(S)$	$3

Picture credits: http://www.toysperiod.com/blog/uncategorized/the-modern-art-and-science-of-origami/

Example 1: Pricing Relational Data

S(Shape,Color,Picture)

Shape	**Color**	**Picture**
Swan	White	
Swan	Yellow	. . . . . .
Dragon	Yellow	
Car	Yellow	. . . . . .
Fish	White	. . . . . .

Price(σ_{Shape})=$2

Price(σ_{Color})=$3

Price list	Price
$V_1 = \sigma_{Shape='Swan'}(S)$	$2
$V_2 = \sigma_{Shape='Dragon'}(S)$	$2
$V_3 = \sigma_{Shape='Car'}(S)$	$2
$V_4 = \sigma_{Shape='Fish'}(S)$	$2
$W_1 = \sigma_{Color='White'}(S)$	$3
$W_2 = \sigma_{Color='Yellow'}(S)$	$3
$W_3 = \sigma_{Color='Red'}(S)$	$3

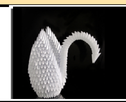
Get all Dragons for $2

Get all Red Origami for $3

Picture credits: http://www.toysperiod.com/blog/uncategorized/the-modern-art-and-science-of-origami/

Example 1: Pricing Relational Data

S(Shape,Color,Picture)

Shape	Color	Picture
Swan	White	
Swan	Yellow	. . . . . .
Dragon	Yellow	
Car	Yellow	. . . . . .
Fish	White	. . . . . .

Price(σ_{Shape})=$2

Price(σ_{Color})=$3

Find the price of the entire db

$1?
$4?
$8?
$20?

Price list	Price
$V_1 = \sigma_{Shape=\text{'Swan'}}(S)$	$2
$V_2 = \sigma_{Shape=\text{'Dragon'}}(S)$	$2
$V_3 = \sigma_{Shape=\text{'Car'}}(S)$	$2
$V_4 = \sigma_{Shape=\text{'Fish'}}(S)$	$2
$W_1 = \sigma_{Color=\text{'White'}}(S)$	$3
$W_2 = \sigma_{Color=\text{'Yellow'}}(S)$	$3
$W_3 = \sigma_{Color=\text{'Red'}}(S)$	$3

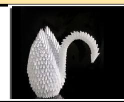
Get all Dragons for $2

Get all Red Origami for $3

Picture credits: http://www.toysperiod.com/blog/uncategorized/the-modern-art-and-science-of-origami/

Example 1: Pricing Relational Data

S(Shape,Color,Picture)

Shape	Color	Picture
Swan	White	
Swan	Yellow	. . . . .
Dragon	Yellow	
Car	Yellow	. . . . .
Fish	White	. . . . .

Price(σ_{Shape})=$2

Price(σ_{Color})=$3

Find the price of the entire db

~~$1?~~
~~$4?~~
$8
~~$20?~~

V_1, V_2, V_3, V_4 determine Q, price(Q) ≤ $8
W_1, W_2, W_3 determine Q, price(Q) ≤ $9

Price list	Price
$V_1 = \sigma_{Shape='Swan'}(S)$	$2
$V_2 = \sigma_{Shape='Dragon'}(S)$	$2
$V_3 = \sigma_{Shape='Car'}(S)$	$2
$V_4 = \sigma_{Shape='Fish'}(S)$	$2
$W_1 = \sigma_{Color='White'}(S)$	$3
$W_2 = \sigma_{Color='Yellow'}(S)$	$3
$W_3 = \sigma_{Color='Red'}(S)$	$3

Get all Dragons for $2

Get all Red Origami for $3

To ensure aribitrage-freeness, we can charge only $8 for the entire database.

Picture credits: http://www.toysperiod.com/blog/uncategorized/the-modern-art-and-science-of-origami/

Example 1: Pricing Relational Data

R

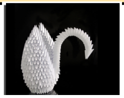
Price(σ_{Shape})=$99

Shape	Instructions
Swan	Fold,fold,fold…
Dragon	Cut,cut,cut,…

S

Price(σ_{Shape})=$2

Price(σ_{Color})=$3

Shape	Color	Picture
Swan	White	
Swan	Yellow	. . . . . .
Dragon	Yellow	
Car	Yellow	. . . . . .
Fish	White	. . . . . .

T

Price(σ_{Color})=$55

Color	PaperSpecs
White	15g/100
Black	20g/100

Find the price of the full join: $Q = R \bowtie S \bowtie T$

Pictures credits: http://www.toysperiod.com/blog/uncategorized/the-modern-art-and-science-of-origami/

Example 1: Pricing Relational Data

R

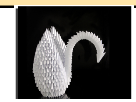
Price(σ_{Shape})=$99

Shape	Instructions
Swan	Fold,fold,fold…
Dragon	Cut,cut,cut,…

S

Price(σ_{Shape})=$2

Price(σ_{Color})=$3

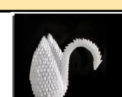
Shape	Color	Picture
Swan	White	
Swan	Yellow	. . . . .
Dragon	Yellow	
Car	Yellow	. . . . .
Fish	White	. . . . .

T

Price(σ_{Color})=$55

Color	PaperSpecs
White	15g/100
Black	20g/100

Find the price of the full join: $Q = R \bowtie S \bowtie T$

Shape	Instructions	Color	Picture	PaperSpecs
Swan	Fold,fold,fold…	White		15g/100

Pictures credits: http://www.toysperiod.com/blog/uncategorized/the-modern-art-and-science-of-origami/

Example 1: Pricing Relational Data

Price(σ_{Shape})=\$99

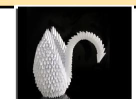
R

Shape	Instructions
Swan	Fold,fold,fold…
Dragon	Cut,cut,cut,…

Price(σ_{Shape})=\$2

Price(σ_{Color})=\$3

S

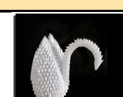
Shape	Color	Picture
Swan	White	
Swan	Yellow	. . . . . .
Dragon	Yellow	
Car	Yellow	. . . . . .
Fish	White	. . . . . .

Price(σ_{Color})=\$55

T

Color	PaperSpecs
White	15g/100
Black	20g/100

Find the price of the full join: $Q = R \bowtie S \bowtie T$

Shape	Instructions	Color	Picture	PaperSpecs
Swan	Fold,fold,fold…	White		15g/100

Not obvious!
E.g. no Yellow Cars in the join.

What to pay for?
$\sigma_{Shape='car'}(R)$ or $\sigma_{Color='yellow'}(T)$

Pictures credits: http://www.toysperiod.com/blog/uncategorized/the-modern-art-and-science-of-origami/

Discussion

Why not charge per row in the answer?

- $Q_1(x,y) = Fortune500(x,y)$
  $Q(x,y) = Fortune500(x,y), StrongBuyRec(x)$
- $Q \subseteq Q_1$, yet $Price(Q) >> Price(Q_1)$
- "Containment" is unrelated to pricing
- "Determinacy" is the right concept for studying pricing

Example 2: Pricing Private Data

UID	User	Rating (0..5)	
1	Alice	3	\$10
2	Bob	0	\$10
3	Carol	1	\$10
4	Dan	0	\$10
…	…	…	
1000	Zoran	2	\$10

- Buyer: query $c = x_1+x_2+\ldots+x_{1000}$
- User compensation: \$10
- Price for the buyer: \$10,000

1. Raw data is too expensive!

Example 2: Pricing Private Data

Differential privacy

- Perturbation is necessary for privacy [Dwork'2011]

Selling private data

- Perturbation is a cost saving feature
- Two extremes:
  - Raw data = no perturbation = high price
  - Differentially private = high perturbation = low price

Example 2: Pricing Private Data

UID	User	Rating (0..5)	
1	Alice	3	$10
2	Bob	0	$10
3	Carol	1	$10
4	Dan	0	$10
…	…	…	
1000	Zoran	2	$10

- Buyer: $c = x_1 + x_2 + \ldots + x_{1000}$
  - Tolerates error ±300
  - Equivalently: variance $v = 5000$*
- Answer: $\hat{c} = c + \mathrm{Lap}(\sqrt{(v/2)})$
- User compensation: ~~$10~~ $0.001 (query is 0.1-DP**)
- Price for the buyer: ~~$10,000~~ $1

2. Perturbation lowers the price

*Probability($|\hat{c} - c| \geq 3\sqrt{2}\,\sigma$) < 1/18=0.056 (Chebyshev), where $\sigma=\sqrt{v} = 50\sqrt{2}$

** $\varepsilon = \sqrt{2}\ \text{sensitivity}(\mathbf{q})/\sigma = 5\sqrt{2} / 50\sqrt{2} = 0.1$

Example 2: Pricing Private Data

UID	User	Rating (0..5)	
1	Alice	3	\$10
2	Bob	0	\$10
3	Carol	1	\$10
4	Dan	0	\$10
…	…	…	
1000	Zoran	2	\$10

- Another buyer: $c = x_1+x_2+\ldots+x_{1000}$
  - ~~Zero error~~, ~~error ±300~~ error ±30
  - ~~Variance = 0, variance = 5000~~ variance = 50
- User compensation: ~~\$10/item,\$0.001/item~~ \$0.1/item? \$1/item?
- Price for the buyer: ~~\$10000, \$1~~ \$100? \$1000?
  - If price > \$100 → arbitrage!
    Buy100 × queries with variance 5000, take average. Cost = 100 × \$1.

3. Multiple queries: must be arbitrage-free.

Outline

- Framework and examples
- Results so far
- Conclusions

Price of Relational Queries

Given: Price points (V_1,p_1), …, (V_k, p_k)
Database D
Arbitrary query Q.

Compute: $\text{Price}_D(Q)$

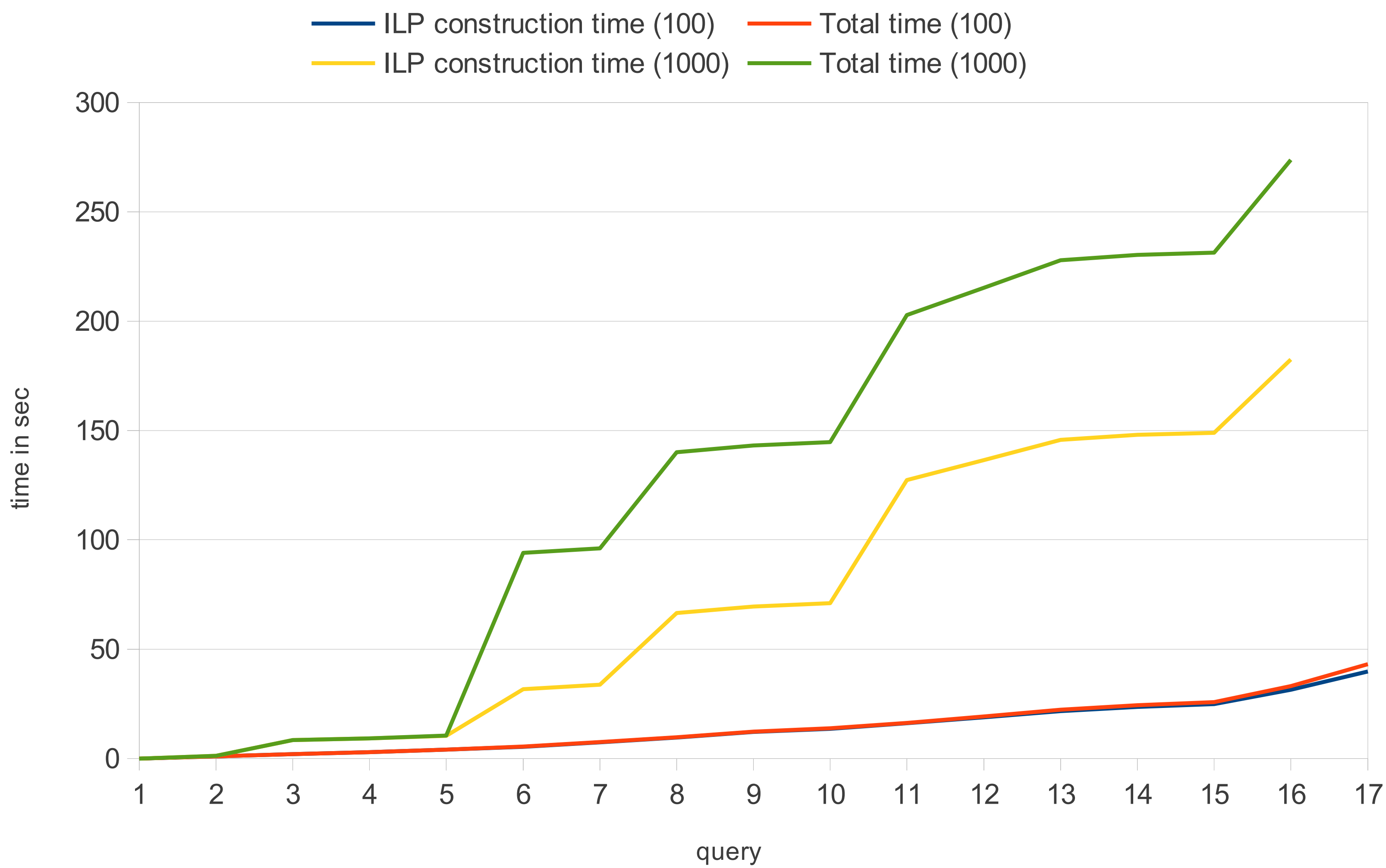
Must ensure this:

Arbitrage-freeness: For all queries,
if Q_1, …, Q_k *determine* Q
then $\text{price}_D(Q) \leq \text{price}_D(Q_1) + \ldots + \text{price}_D(Q_k)$

Price of Relational Queries

- Simple algorithm for computing $price_D(Q)$ given an oracle for checking *deteminacy*
- Two options for *determinacy*
  - *Instance-independent*: used by RDBMS today in query-answering using views; undecidable!
  - *Instance-dependent*: seems more natural for pricing; Π^p_2 in the database
- If (a) price-points (V_i,p_i) are selection queries, and (b) Q is a Union of Conjunctive Queries then $price_D(Q)$ is NP-complete in the database
- Reduction to ILP makes pricing (almost) practical

Price of Relational Queries



Compensation for Private Data

UID	User	Rating (0..5)	
1	Alice	3	\$10
2	Bob	0	\$10
3	Carol	1	\$10
4	Dan	0	\$10
…	…	…	
1000	Zoran	2	\$10

Query $c = x_1+x_2+\ldots+x_{1000}$
Variance $v = 50$

How much should we pay Carol?

Compensation for Private Data

UID	User	Rating (0..5)	
1	Alice	3	\$10
2	Bob	0	\$10
3	Carol	1	\$10
4	Dan	0	\$10
…	…	…	
1000	Zoran	2	\$10

Query $c = x_1+x_2+\ldots+x_{1000}$
Variance $v = 50$

How much should we pay Carol?

Differential Privacy

Def. [Dwork'11] Fix ε. Mechanism $\hat{c}$ is called ε-differential private, if for all D, D' that differ in one item, and any set S

$$P[\hat{c}(D) \in S] \leq \exp(\varepsilon) \times P[\hat{c}(D') \in S]$$

Compensation for Private Data

UID	User	Rating (0..5)	
1	Alice	3	\$10
2	Bob	0	\$10
3	Carol	1	\$10
4	Dan	0	\$10
…	…	…	
1000	Zoran	2	\$10

Query $c = x_1+x_2+\ldots+x_{1000}$
Variance $v = 50$

How much should we pay Carol?

Differential Privacy

Def. [Dwork'11] Fix ε. Mechanism $\hat{c}$ is called ε-differential private, if for all D, D' that differ in one item, and any set S

$$P[\hat{c}(D) \in S] \leq \exp(\varepsilon) \times P[\hat{c}(D') \in S]$$

Thm. The mechanism $\hat{c}(D) = c(D) + \mathrm{Lap}(\Delta c/\varepsilon)$ is ε-differential private

Variance $v=2(\Delta c/\varepsilon)^2$

Carol gets no money!

Compensation for Private Data

UID	User	Rating (0..5)	
1	Alice	3	\$10
2	Bob	0	\$10
3	Carol	1	\$10
4	Dan	0	\$10
…	…	…	
1000	Zoran	2	\$10

Query $c = x_1+x_2+\ldots+x_{1000}$
Variance $v = 50$

How much should we pay Carol?

Differential Privacy

Def. [Dwork'11] Fix ε. Mechanism $\hat{c}$ is called ε-differential private, if for all D, D' that differ in one item, and any set S

$$P[\hat{c}(D) \in S] \le \exp(\varepsilon) \times P[\hat{c}(D') \in S]$$

Thm. The mechanism $\hat{c}(D) = c(D) + \mathrm{Lap}(\Delta c/\varepsilon)$ is ε-differential private

Variance $v=2(\Delta c/\varepsilon)^2$

Carol gets no money!

Data Pricing

Fix variance v

Carol's compensation W depends on ε which depends on v

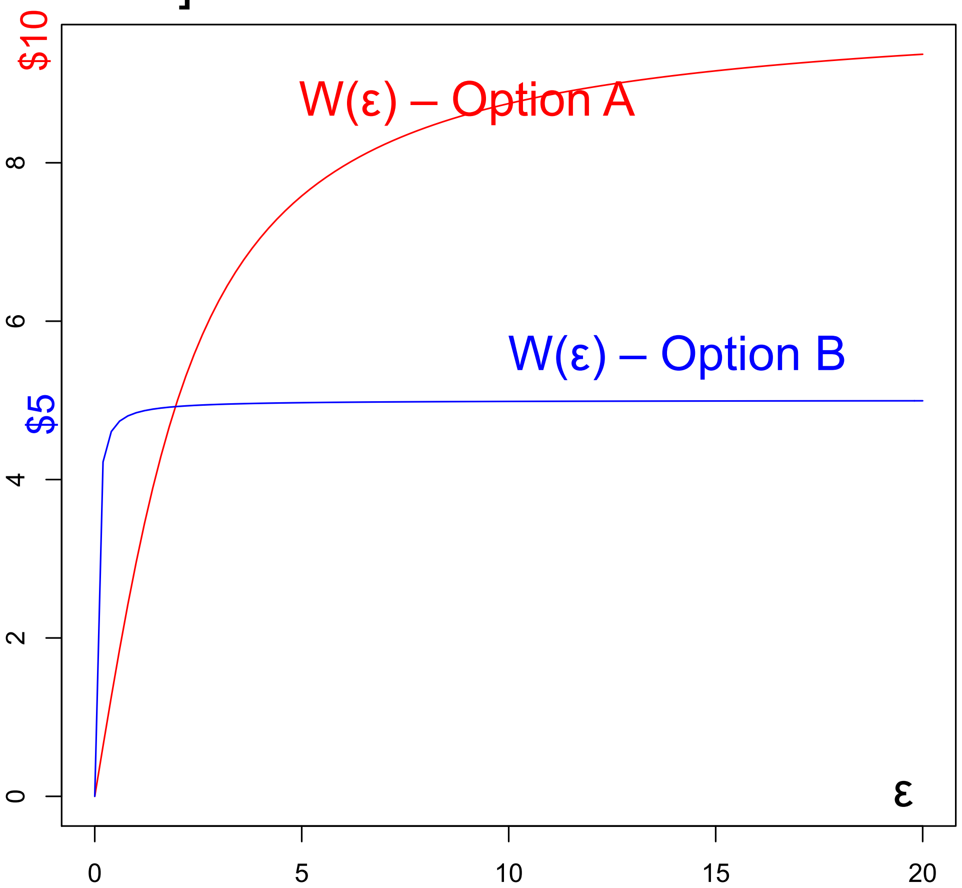
Def. Carol's privacy loss is
$\varepsilon(v) = \sup_S \log(P[\hat{c}(D) \in S]/P[\hat{c}(D') \in S])$

$W(\varepsilon)$ = Carol's valuation function

Compensation for Private Data

Incentivizing Carol to reveal her valuation W(ε) is difficult!
[Ghosh'11,Gkatzelis'12,Riederer'12]
We use an idea from [Aperjis&Huberman'11]:

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

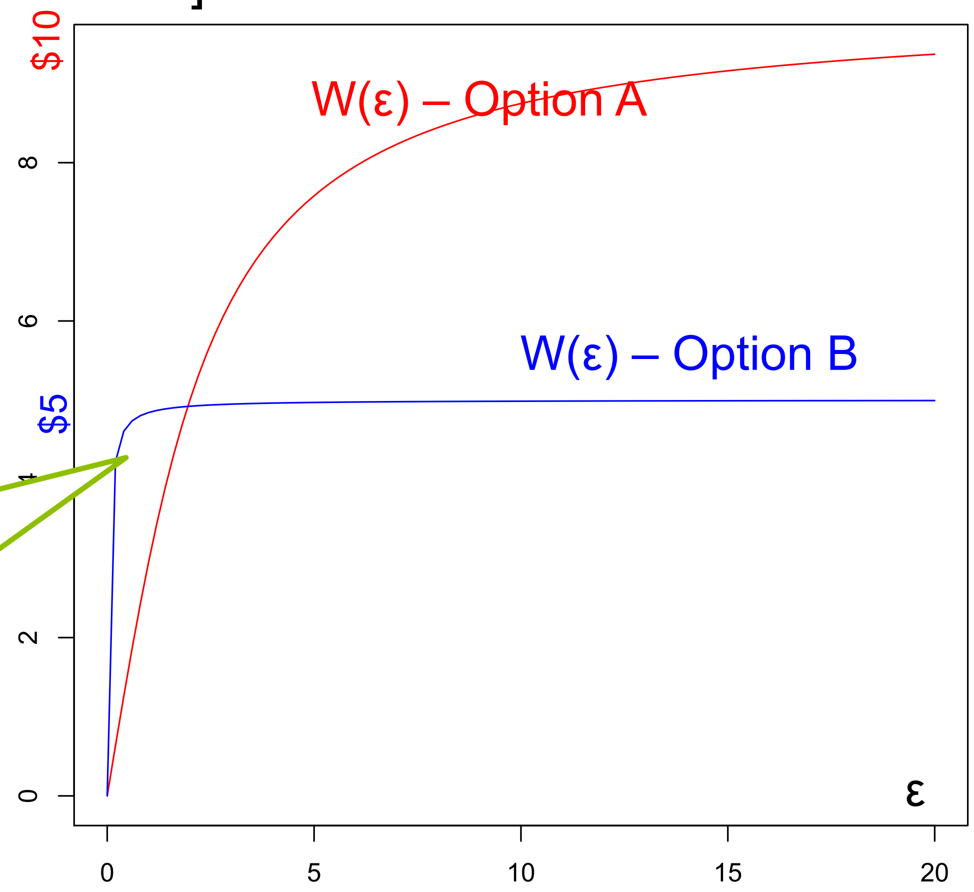


Compensation for Private Data

Incentivizing Carol to reveal her valuation W(ε) is difficult! [Ghosh'11,Gkatzelis'12,Riederer'12]

We use an idea from [Aperjis&Huberman'11]:

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

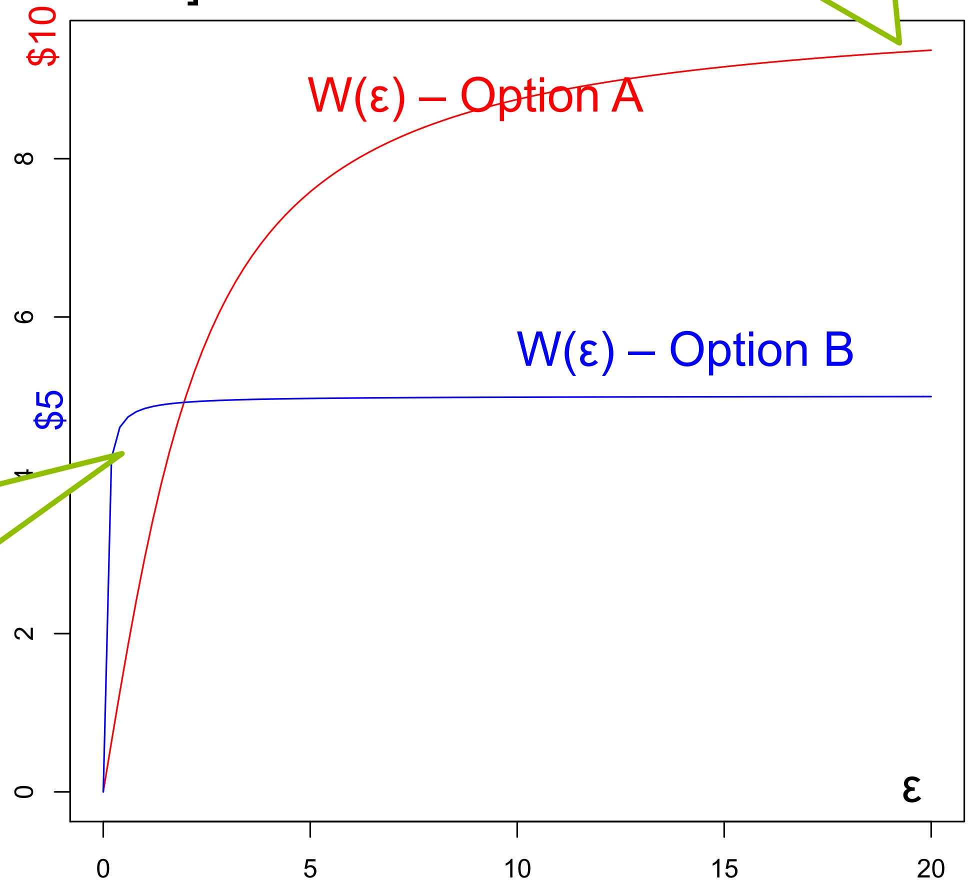


Risk-averse users count on the fact that most queries will have low privacy leak

Compensation for Priva

Incentivizing Carol to reveal her valuation W(ε) is difficu
[Ghosh'11,Gkatzelis'12,Riederer'12]
We use an idea from [Aperjis&Huberman'11]:

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



Risk-neutral users want full compensation at the risk of never being paid

Risk-averse users count on the fact that most queries will have low privacy leak

Outline

- Framework and examples
- Results so far
- Conclusions

The Third Wave of Computing

- First wave = hardware
  - IBM, DEC, Sun, …
  - 1950 – 1980
- Second wave = software
  - Microsoft, Borland, Fox Software, Oracle, …
  - 1980 -- 2010
- Third wave = data!
  - Google maps v.s. IOS maps
  - Facebook’s users

Conclusions

- Data has (lots of) value!
- Pricing data: at the intersection of three areas:
  - Data management
  - Mechanism design
  - Economics

This talk

- Key concepts:
  - Arbitrage-free
  - Compensation = function of privacy loss

References

- Koutris et al., PODS, 2012
- Li et al., ICDT, 2013
- Koutris et al, under review