

# 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

# AggData

## Complete List of Shell Gas Stations

This is a complete list of Shell gas station locations including geographic coordinates. Shell gas stations carry fuel as well as a variety of car care products.

### AggDATA PROFILE

Number of Records: 14185  
Last Updated: 03/25/2011  
File Format: **CSV**  
Regions Included: US only  
Categories: **Gas Stations**

### Fields Included:

- Address
- City
- State
- Zip Code
- Phone Number
- Latitude
- Longitude

Download  
a Sample  
of this  
Aggdata



### Local Gas Stations

Find Local Gas Station Listings - 24 Hour, Full Service, Car Washes!  
[yellowpages.com](http://yellowpages.com)

Ads by Google

## Purchase AggData

ADD TO CART

VIEW CART



\$59.00

Can't find the data you're looking for? Request a custom list!

### Find AggData:

Just start typing the name of the company or category.

Search

### Related Data

[Complete List of Holiday Station Locations](#)

[Complete List of 7-11 Canada Locations](#)

[Complete List of Tedeschi Food Shops Locations](#)

[Complete List of Marathon Gas Locations](#)

[Complete List of Thorntons Locations](#)

ADD THIS



\$699 for 885976 teacher names & emails!

CALL US ON 1800 495 9313

CustomLists.net  
DATABASES FOR YOUR BUSINESS



Databases Available Lists Guarantee FAQ About Us Contact Us Your Shopping Cart

- American
  - Residential Schools
- Canadian
  - Business
  - Consumer
  - Residential Schools
- Australian
  - Business
  - Consumer
  - Residential Schools
- New Zealand
  - Business
  - Consumer
  - Residential

### American Teacher Email Database

Last Updated December 2012

- 885,976 Listings
- 885,976 Email Addresses
- 0 Phone Numbers
- 0 Fax Numbers
- 0 Addresses

Normally: USD\$1,299.00  
Special Price: USD\$699.00

BUY NOW

#### All Data Verified & Cross Checked via:

- Individual Businesses
- ABN Records
- ABN Records
- ATO Records

#### Listings Include:

- ✓ Email Address

#### All Data Verified and Cross-Checked via:

- ✓ Department of Education
- ✓ Yellow Pages Directory
- ✓ White Pages Directory

#### The Most Comprehensive Database of American Schools:

- ✓ 885,976 Listings
- ✓ 885,976 Email Addresses

Normally: USD\$1,299.00  
Special Price USD\$699.00

BUY NOW

Download a FREE Sample List  
Find out just how comprehensive and easy to use our lists are with these samples of the full database.

Your Email Address

Select Your Database

Get Your Free Database

#### Why Custom Lists:

- Cost effective marketing
- Regularly Updated Databases
- Accurate & Easy To Use
- Download Instantly



100% Satisfaction Guarantee



Cheaper just for Washington

CustomLists.net  
MARKETING DATABASES FOR YOUR BUSINESS

CALL US ON 1800 495 9313

Need Help? Have Questions? OK, Let's Chat

SECURED BY RapidSSL 2048-bit root

Home Available Databases Available Lists Guarantee FAQ About Us Contact Us Your Shopping Cart

United Kingdom  
Business  
Consumer  
Residential  
Schools

American  
Business  
Consumer  
Residential  
Schools

Canadian  
Business  
Consumer  
Residential  
Schools

Australian  
Business  
Consumer  
Residential  
Schools

New Zealand  
Business  
Consumer  
Residential  
Schools

**Washington Schools Database**  
Updated December 2012

- 2,926 Listings
- 0 Email Addresses
- 306 Phone Numbers
- 0 Fax Numbers
- 2,926 Addresses

Normally: ~~USD\$449.00~~  
**Special Price: USD\$249.00**

» BUY NOW

**All Data Verified & Cross Checked via:**

- Individual Businesses
- ABN Records
- ABN Records
- ATO Records

**Listings Include:**

- ✓ School Name
- ✓ Address
- ✓ Phone Number
- ✓ Year Level
- ✓ Enrollments
- ✓ Teachers
- ✓ Ethnicity

**All Data Verified and Cross-Checked via:**

- ✓ Department of Education
- ✓ Yellow Pages Directory
- ✓ White Pages Directory

**The Most Comprehensive Database of American Schools:**

- ✓ 2,926 Listings
- ✓ 306 Phone Numbers
- ✓ 2,926 Addresses

Normally: ~~USD\$449.00~~  
**Special Price USD\$249.00**

» BUY NOW

**Download a FREE Sample List**  
Find out just how comprehensive and easy to use our lists are with these samples of the full database.

Your Email Address

Select Your Database

Get Your Free Database

**Why Custom Lists:**

- Cost effective marketing
- Regularly Updated Databases
- Accurate & Easy To Use
- Download Instantly

**100% SATISFACTION GUARANTEED**

100% Satisfaction Guarantee

SECURED BY RapidSSL 2048-bit root

**What People Are Saying About Us:**

"We marketed our business seminar using the United Kingdom Business Database. We achieved an open rate of 77% and 236 registrations in the first day!"

# 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:  
 $(V_1, p_1), (V_2, p_2), \dots$  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  $Q_1, \dots, Q_k, Q$ ,

if  $Q_1, \dots, Q_k$  determine  $Q$ , then:

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

“ $Q_1, \dots, Q_k$  *determine*  $Q$ ” means that

$Q(D)$  can be answered from  $Q_1(D), \dots, Q_k(D)$ ,

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( $\sigma_{\text{Shape}}$ )=\$2

Price( $\sigma_{\text{Color}}$ )=\$3

Price list

Price

$$V_1 = \sigma_{\text{Shape}='Swan'}(S) \quad \$2$$

$$V_2 = \sigma_{\text{Shape}='Dragon'}(S) \quad \$2$$

$$V_3 = \sigma_{\text{Shape}='Car'}(S) \quad \$2$$

$$V_4 = \sigma_{\text{Shape}='Fish'}(S) \quad \$2$$

$$W_1 = \sigma_{\text{Color}='White'}(S) \quad \$3$$

$$W_2 = \sigma_{\text{Color}='Yellow'}(S) \quad \$3$$

$$W_3 = \sigma_{\text{Color}='Red'}(S) \quad \$3$$

# Example 1: Pricing Relational Data

S(Shape,Color,Picture)

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

Price( $\sigma_{\text{Shape}}$ )=\$2

Price( $\sigma_{\text{Color}}$ )=\$3

Price list

Price

$$V_1 = \sigma_{\text{Shape}='Swan'}(S) \quad \$2$$

$$V_2 = \sigma_{\text{Shape}='Dragon'}(S) \quad \$2$$

$$V_3 = \sigma_{\text{Shape}='Car'}(S) \quad \$2$$

$$V_4 = \sigma_{\text{Shape}='Fish'}(S) \quad \$2$$

$$W_1 = \sigma_{\text{Color}='White'}(S) \quad \$3$$

$$W_2 = \sigma_{\text{Color}='Yellow'}(S) \quad \$3$$

$$W_3 = \sigma_{\text{Color}='Red'}(S) \quad \$3$$

Get all Dragons for \$2

Get all Red Origami for \$3

# Example 1: Pricing Relational Data

S(Shape,Color,Picture)

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

Price( $\sigma_{\text{Shape}}$ )=\$2

Price( $\sigma_{\text{Color}}$ )=\$3

Find the price of the entire db

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

Price list

Price

$V_1 = \sigma_{\text{Shape}='Swan'}(S)$      \$2

$V_2 = \sigma_{\text{Shape}='Dragon'}(S)$      \$2

$V_3 = \sigma_{\text{Shape}='Car'}(S)$      \$2

$V_4 = \sigma_{\text{Shape}='Fish'}(S)$      \$2

$W_1 = \sigma_{\text{Color}='White'}(S)$      \$3

$W_2 = \sigma_{\text{Color}='Yellow'}(S)$      \$3

$W_3 = \sigma_{\text{Color}='Red'}(S)$      \$3

Get all Dragons for \$2

Get all Red Origami for \$3

# Example 1: Pricing Relational Data

S(Shape,Color,Picture)

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

Price( $\sigma_{\text{Shape}}$ )=\$2

Price( $\sigma_{\text{Color}}$ )=\$3

Find the price of the entire db

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

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

Price list

Price

$V_1 = \sigma_{\text{Shape}='Swan'}(S)$      \$2

$V_2 = \sigma_{\text{Shape}='Dragon'}(S)$      \$2

$V_3 = \sigma_{\text{Shape}='Car'}(S)$      \$2

$V_4 = \sigma_{\text{Shape}='Fish'}(S)$      \$2

$W_1 = \sigma_{\text{Color}='White'}(S)$      \$3

$W_2 = \sigma_{\text{Color}='Yellow'}(S)$      \$3

$W_3 = \sigma_{\text{Color}='Red'}(S)$      \$3

Get all Dragons for \$2

Get all Red Origami for \$3

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

# Example 1: Pricing Relational Data

R

Price( $\sigma_{\text{Shape}}$ )=\$99

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

S

Price( $\sigma_{\text{Shape}}$ )=\$2

Price( $\sigma_{\text{Color}}$ )=\$3

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

T

Price( $\sigma_{\text{Color}}$ )=\$55

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

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

# Example 1: Pricing Relational Data

R

Price( $\sigma_{\text{Shape}}$ )=\$99

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

S

Price( $\sigma_{\text{Shape}}$ )=\$2

Price( $\sigma_{\text{Color}}$ )=\$3

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

T

Price( $\sigma_{\text{Color}}$ )=\$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

# Example 1: Pricing Relational Data

R

Price( $\sigma_{\text{Shape}}$ )=\$99

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

S

Price( $\sigma_{\text{Shape}}$ )=\$2

Price( $\sigma_{\text{Color}}$ )=\$3

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

T

Price( $\sigma_{\text{Color}}$ )=\$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

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

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

# Discussion

Why not charge per row in the answer?

- $Q_1(x,y) = \text{Fortune500}(x,y)$   
 $Q(x,y) = \text{Fortune500}(x,y), \text{StrongBuyRec}(x)$
- $Q \subseteq Q_1$ , yet  $\text{Price}(Q) \gg \text{Price}(Q_1)$
- “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  $c = x_1 + x_2 + \dots + x_{1000}$
- 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:  $c = x_1 + x_2 + \dots + x_{1000}$ 
  - Tolerates error  $\pm 300$
  - Equivalently: variance  $v = 5000^*$
- Answer:  $\hat{c} = c + \text{Lap}(\sqrt{(v/2)})$
- 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  $\sigma = \sqrt{v} = 50\sqrt{2}$

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

# 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:  $c = x_1 + x_2 + \dots + x_{1000}$ 
  - ~~Zero error, error  $\pm 300$~~  error  $\pm 30$
  - ~~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  $\rightarrow$  arbitrage!  
Buy 100  $\times$  queries with variance 5000, take average. Cost = 100  $\times$  \$1.

3. Multiple queries: must be arbitrage-free.

# Outline

- Framework and examples
- Results so far
- Conclusions

# Price of Relational Queries

**Given**: Price points  $(V_1, p_1), \dots, (V_k, p_k)$

Database  $D$

Arbitrary query  $Q$ .

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

Must ensure this:

**Arbitrage-freeness**: For all queries,

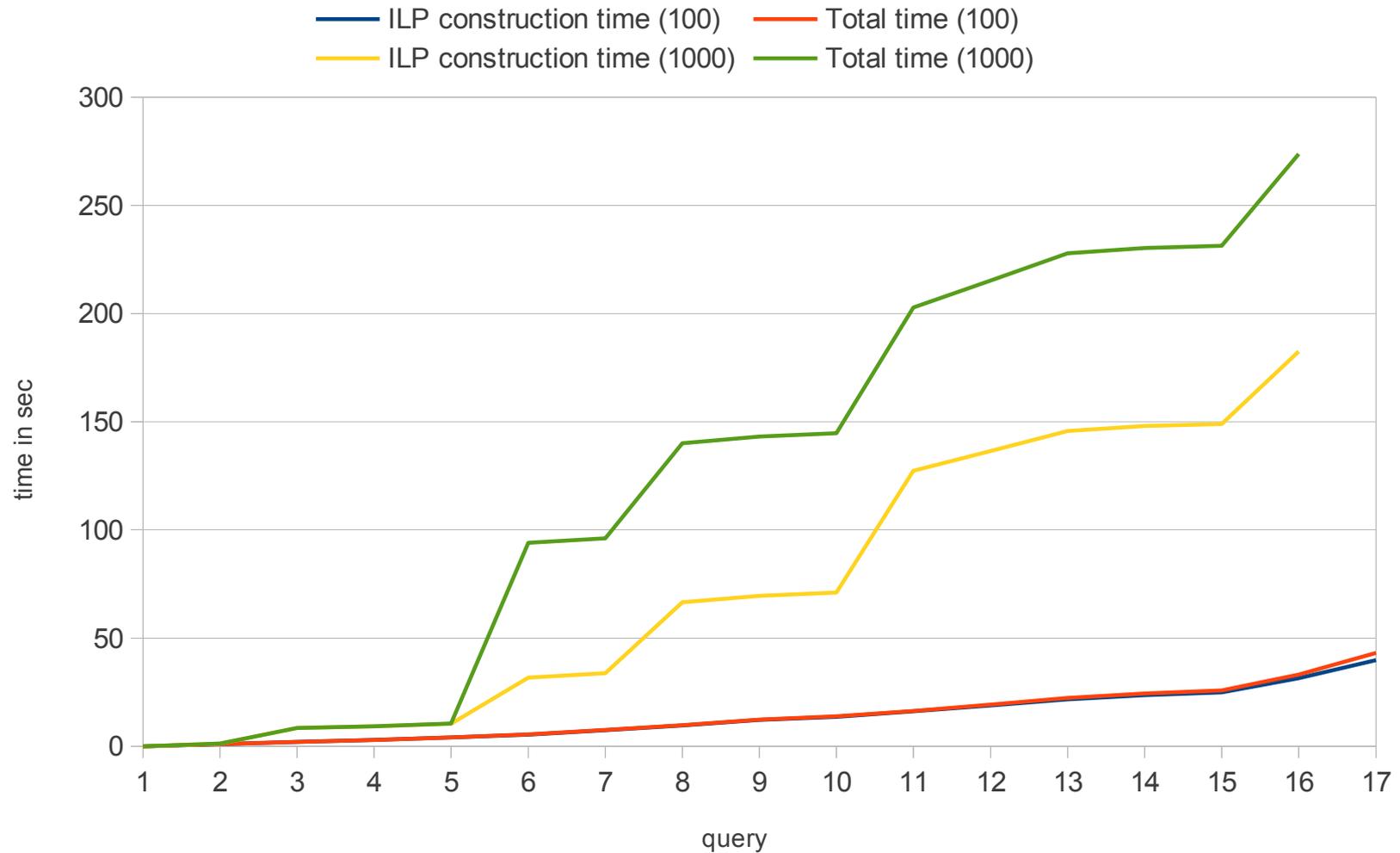
if  $Q_1, \dots, Q_k$  determine  $Q$

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

# Price of Relational Queries

- **Simple algorithm** for computing  $\text{price}_D(Q)$  given an oracle for checking determinacy
- **Two options** for determinacy
  - Instance-independent: used by RDBMS today in query-answering using views; **undecidable!**
  - Instance-dependent: seems more natural for pricing;  $\Pi^P_2$  in the database
- If (a) price-points  $(V_i, p_i)$  are **selection queries**, and (b)  $Q$  is a **Union of Conjunctive Queries** then  $\text{price}_D(Q)$  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  $c = x_1 + x_2 + \dots + x_{1000}$

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  $c = x_1 + x_2 + \dots + x_{1000}$

Variance  $v = 50$

How much should we pay Carol?

## Differential Privacy

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

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

# 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  $c = x_1 + x_2 + \dots + x_{1000}$   
Variance  $v = 50$

How much should we pay Carol?

## Differential Privacy

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

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

**Thm.** The mechanism  $\hat{c}(D) = c(D) + \text{Lap}(\Delta c / \epsilon)$  is  $\epsilon$ -differential private

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

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

## Differential Privacy

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

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

**Thm.** The mechanism  $\hat{c}(D) = c(D) + \text{Lap}(\Delta c/\epsilon)$  is  $\epsilon$ -differential private

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

Carol gets no money!

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

How much should we pay Carol?

## Data Pricing

Fix variance  $v$

Carol's compensation  $W$  depends on  $\epsilon$  which depends on  $v$

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

$W(\epsilon) = \text{Carol's valuation function}$

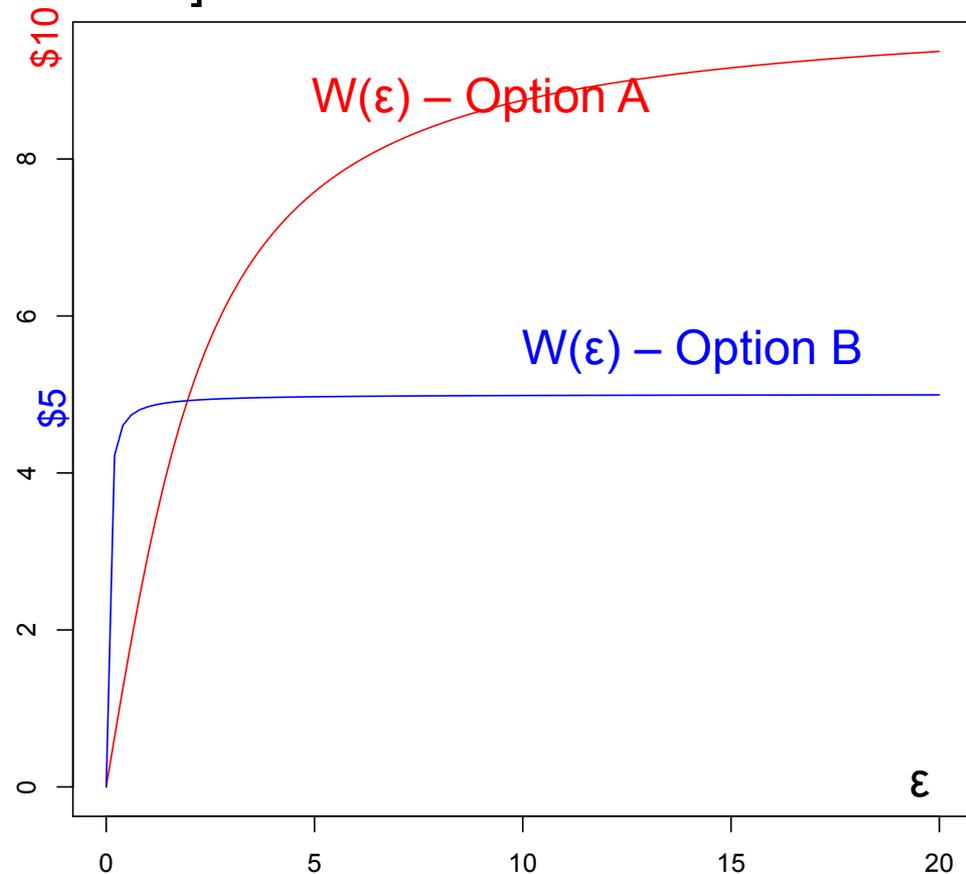
# Compensation for Private Data

Incentivizing Carol to reveal her valuation  $W(\varepsilon)$  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

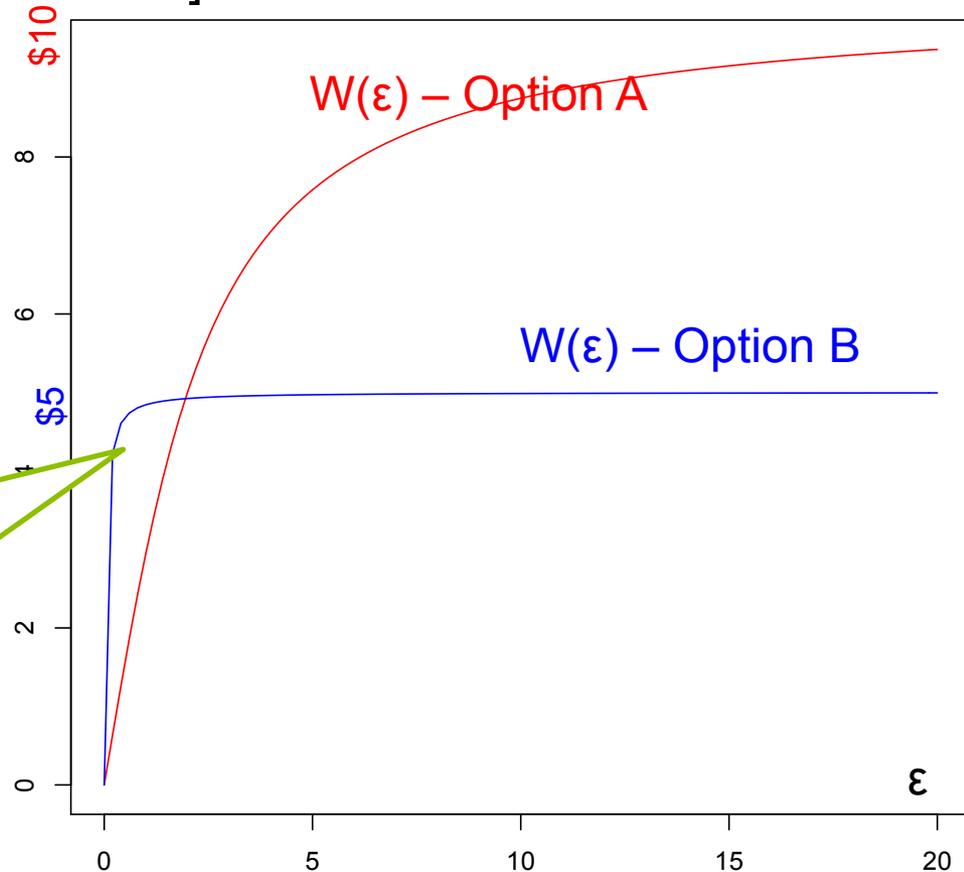
Incentivizing Carol to reveal her valuation  $W(\epsilon)$  is difficult!

[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 Privacy

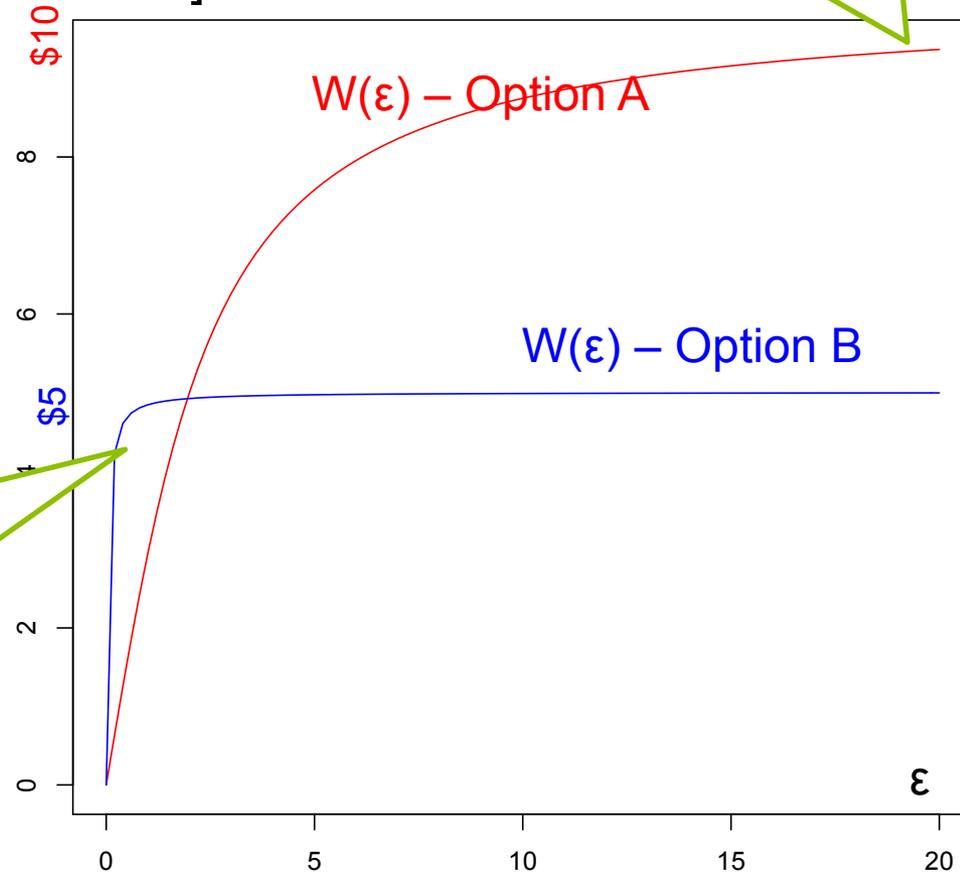
Incentivizing Carol to reveal her valuation  $W(\epsilon)$  is difficult  
[Ghosh'11, Gkatzelis'12, Riederer'12]

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

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

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



This talk

# References

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