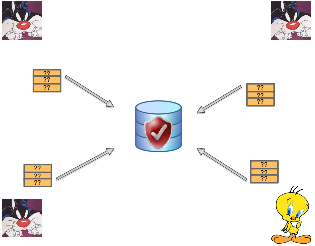
Private Equilibrium Computation for Analyst Privacy

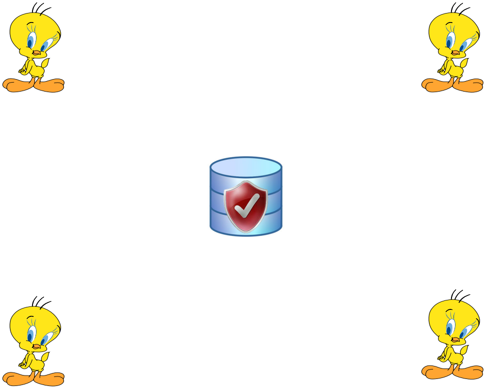


Justin Hsu, Aaron Roth,[1] Jonathan Ullman[2]

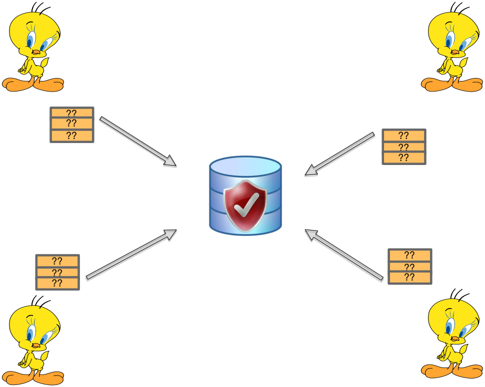
[1]University of Pennsylvania
[2]Harvard University

June 2, 2013

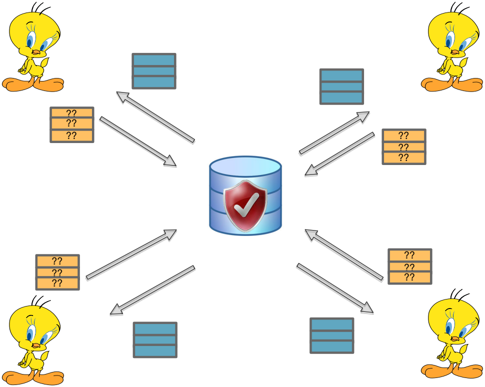
A market survey scenario

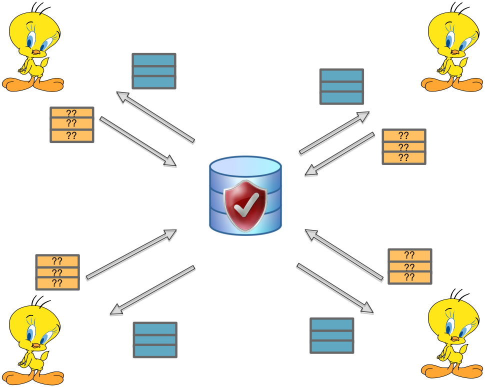


A market survey scenario



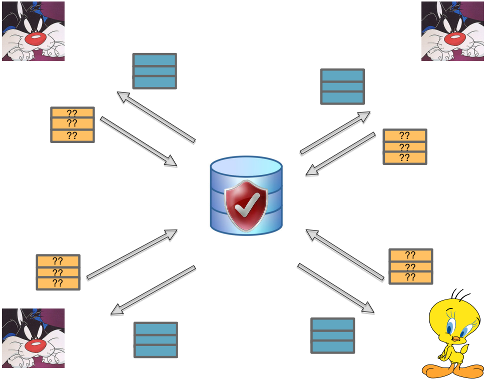
A market survey scenario





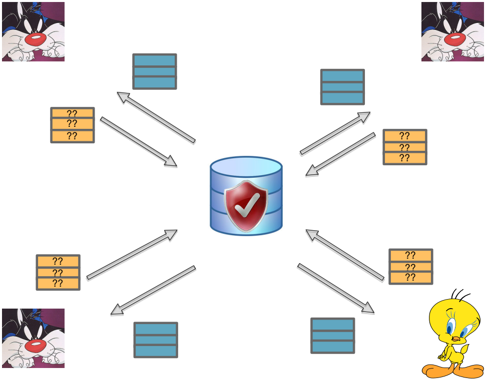
Requirements

- Data privacy: protect the consumer's privacy



Requirements

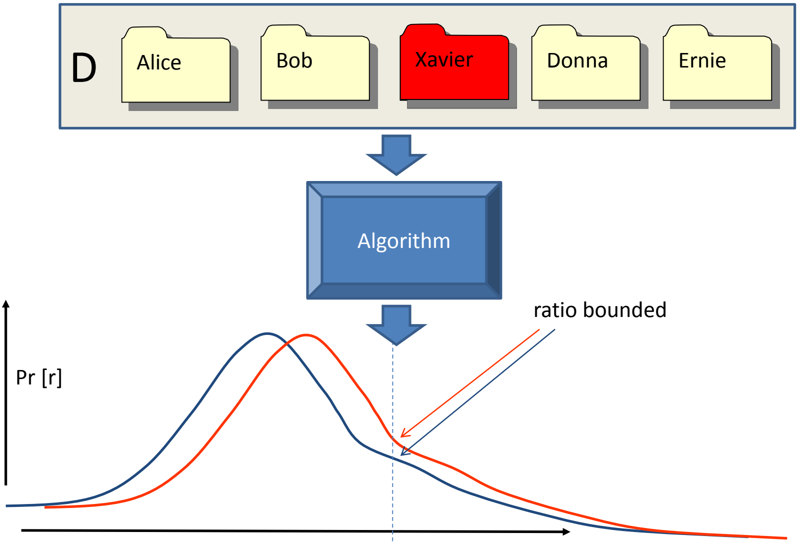
- Data privacy: protect the consumer's privacy



Requirements

- Data privacy: protect the consumer's privacy
- Analyst privacy [DNV'12]: protect the analyst's privacy

(Standard) Differential privacy [DMNS'06]



Definition (DMNS'06)

Let M be a randomized mechanism from databases to range $\mathcal{R}$, and let D, D' be databases differing in one record. M is ϵ-differentially private if for every $r \in \mathcal{R}$,

$$\Pr[M(D) = r] \le e^{\epsilon} \cdot \Pr[M(D') = r].$$

Useful properties

- Very strong, worst-case privacy guarantee
- Well-behaved under composition, post-processing

Many-to-one-analyst privacy [DNV'12]

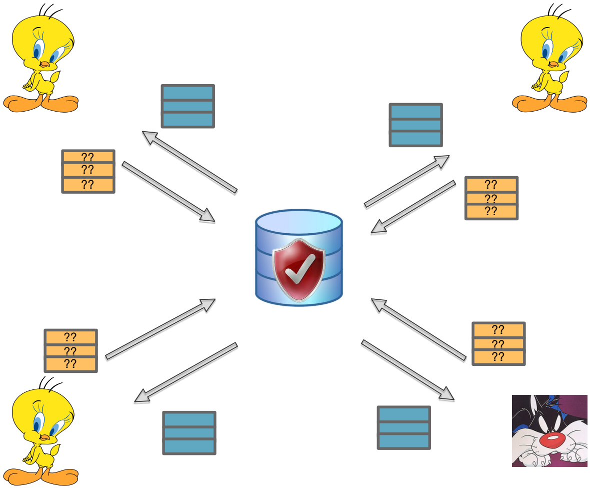
Intuition

- A single analyst can't tell if other analysts change their queries

Many-to-one-analyst privacy [DNV'12]

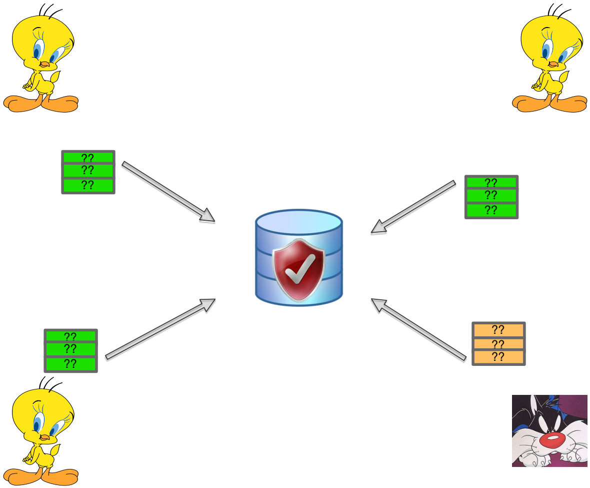
Intuition

- A single analyst can't tell if other analysts change their queries



Intuition

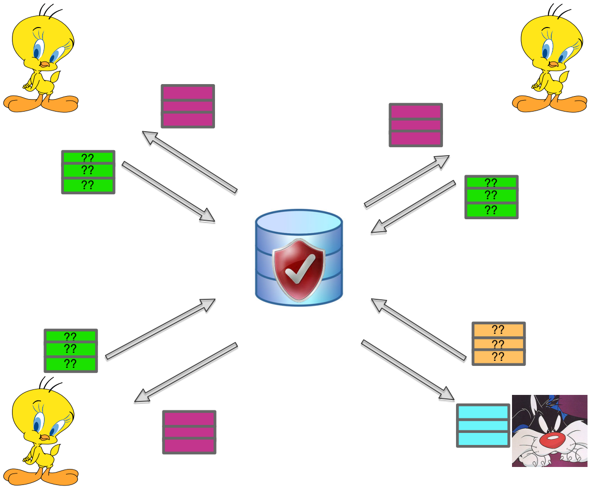
- A single analyst can't tell if other analysts change their queries



Many-to-one-analyst privacy [DNV'12]

Intuition

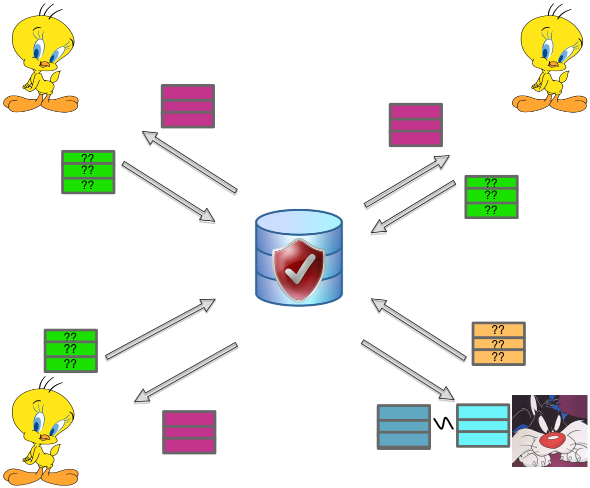
- A single analyst can't tell if other analysts change their queries



Many-to-one-analyst privacy [DNV'12]

Intuition

- A single analyst can't tell if other analysts change their queries



One-query-to-many-analyst privacy (Today)

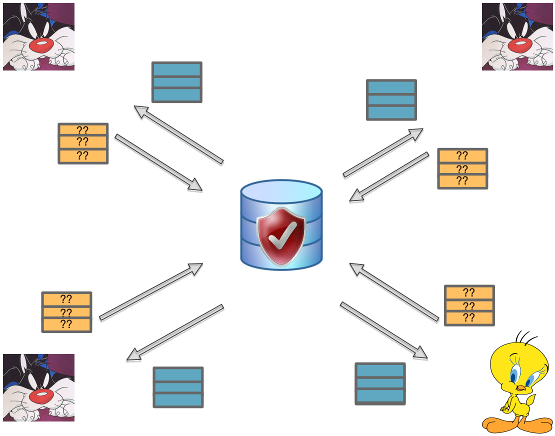
Intuition

- All but one analyst (possibly colluding) can't tell if last analyst changes one of their queries

One-query-to-many-analyst privacy (Today)

Intuition

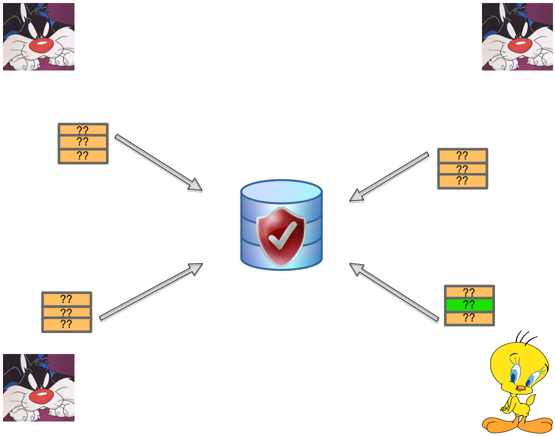
- All but one analyst (possibly colluding) can't tell if last analyst changes one of their queries



One-query-to-many-analyst privacy (Today)

Intuition

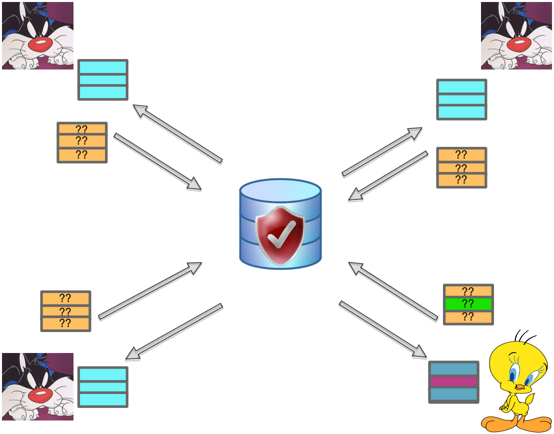
- All but one analyst (possibly colluding) can't tell if last analyst changes one of their queries



One-query-to-many-analyst privacy (Today)

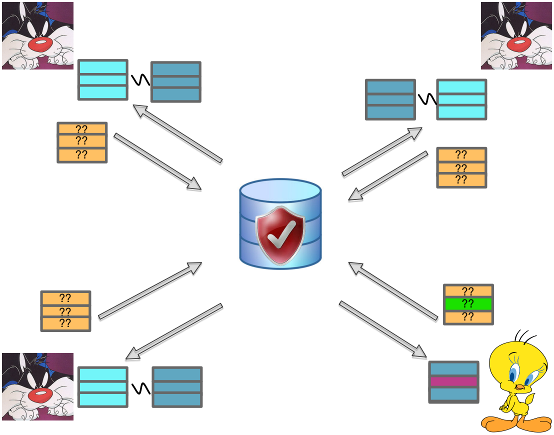
Intuition

- All but one analyst (possibly colluding) can't tell if last analyst changes one of their queries



Intuition

- All but one analyst (possibly colluding) can't tell if last analyst changes one of their queries



Basic problem

- Analysts want accurate answers to a large set $\mathcal{Q}$ of counting (linear) queries

Basic problem

- Analysts want accurate answers to a large set $\mathcal{Q}$ of counting (linear) queries

"What fraction of records satisfy P?"

Basic problem

- Analysts want accurate answers to a large set $\mathcal{Q}$ of counting (linear) queries

"What fraction of records satisfy P?"

- Privately construct synthetic database to answer queries

Basic problem

- Analysts want accurate answers to a large set $\mathcal{Q}$ of counting (linear) queries

"What fraction of records satisfy P?"

- Privately construct synthetic database to answer queries

Prior work

- Long line of work [BLR'08, RR'09, HR'10,...], data privacy

Basic problem

- Analysts want accurate answers to a large set $\mathcal{Q}$ of counting (linear) queries

  "What fraction of records satisfy P?"

- Privately construct synthetic database to answer queries

Prior work

- Long line of work [BLR'08, RR'09, HR'10,. . .], data privacy
- Stateful mechanisms: not analyst private

Theorem

Suppose the analysts ask queries $\mathcal{Q}$, and let the database have n records from $\mathcal{X}$. There exists an ϵ analyst and data private mechanism which achieves error α on all queries in $\mathcal{Q}$, where

$$\alpha = O\left(\frac{\mathrm{polylog}(|\mathcal{X}|, |\mathcal{Q}|)}{\epsilon\sqrt{n}}\right).$$

Plan for rest of the talk

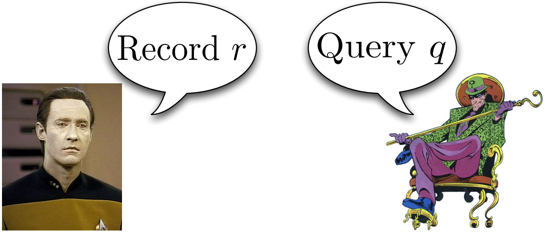
Outline

- Interpretation of query release as a game
- Privately solving the query release game
- Analyst private query release

The query release game

The query release game

The query release game



The query release game

Record r

Query q

Loss $q(r) - q(D)$

(D is true database)

Record r

Query q

Loss $q(r) - q(D)$

Loss $-(q(r) - q(D))$

(D is true database)

Database as a distribution

- Think of true database D as a distribution over records
- $\hat{D}$ is data player's distribution over records

Database as a distribution

- Think of true database D as a distribution over records
- $\hat{D}$ is data player's distribution over records

Mixed strategy

Database as a distribution

- Think of true database D as a distribution over records
- $\hat{D}$ is data player's distribution over records

Mixed strategy

- Versus a counting query q, data player's expected loss:

$$\mathbb{E}_{r\sim\hat{D}}[q(r) - q(D)] = q(\hat{D}) - q(D)$$

Database as a distribution

- Think of true database D as a distribution over records
- $\hat{D}$ is data player's distribution over records

Mixed strategy

- Versus a counting query q, data player's expected loss:

$$\mathbb{E}_{r \sim \hat{D}}[q(r) - q(D)] = q(\hat{D}) - q(D)$$

- D is mixed strategy with zero loss

Equilibrium strategy

What if small expected loss?

- Suppose data player's <u>expected loss less than α</u> for all queries

α-approximate equilibrium

What if small expected loss?

- Suppose data player's expected loss less than α for all queries

α-approximate equilibrium

What if small expected loss?

- Suppose data player's expected loss less than α for all queries
- Data distribution answers all queries with error at most α

α-approximate equilibrium

What if small expected loss?

- Suppose data player's expected loss less than α for all queries
- Data distribution answers all queries with error at most α

Query release!

α-approximate equilibrium

What if small expected loss?

- Suppose data player's expected loss less than α for all queries
- Data distribution answers all queries with error at most α

Synthetic database

Query release!

What if small expected loss?

α-approximate equilibrium

- Suppose data player's expected loss less than α for all queries
- Data distribution answers all queries with error at most α

Synthetic database

Query release!

- But how to compute this?

Computing the equilibrium ~~privately~~

Known approach: repeated game

- Players maintain distributions over actions

Known approach: repeated game

- Players maintain distributions over actions
- Loop:
  - Sample and play action

Computing the equilibrium ~~privately~~

Known approach: repeated game

- Players maintain distributions over actions
- Loop:
  - Sample and play action
  - Receive loss for all actions

Computing the equilibrium ~~privately~~

Known approach: repeated game

- Players maintain distributions over actions
- Loop:
  - Sample and play action
  - Receive loss for all actions
  - Update distribution: increase probability of better actions

Known approach: repeated game

- Players maintain distributions over actions
- Loop:
  - Sample and play action
  - Receive loss for all actions
  - Update distribution: increase probability of better actions

Multiplicative weights (MW)

Computing equilibrium strategy ~~privately~~

Record r

Query q

Loss $q(r) - q(D)$

Loss $-(q(r) - q(D))$

Computing equilibrium strategy ~~privately~~

MW

MW

Loss $q(r) - q(D)$

Loss $-(q(r) - q(D))$

Computing equilibrium strategy ~~privately~~

Computing equilibrium strategy privately

Idea: use distribution over plays [FS'96]

- Both players use multiplicative weights

- MW distributions converge to approximate equilibrium

Idea: use distribution over plays [FS'96]

- Both players use multiplicative weights

Not private

- ~~MW distributions converge to approximate equilibrium~~

Computing equilibrium strategy privately

Idea: use distribution over plays [FS'96]

- Both players use multiplicative weights

Not private

- ~~MW distributions converge to approximate equilibrium~~
- Empirical distributions also converge to approximate equilibrium

Computing equilibrium strategy privately

Idea: use distribution over plays [FS'96]

- Both players use multiplicative weights

Not private

- ~~MW distributions converge to approximate equilibrium~~
- Empirical distributions also converge to approximate equilibrium

Distribution of actual plays

Computing equilibrium strategy privately

Idea: use distribution over plays [FS'96]

- Both players use multiplicative weights

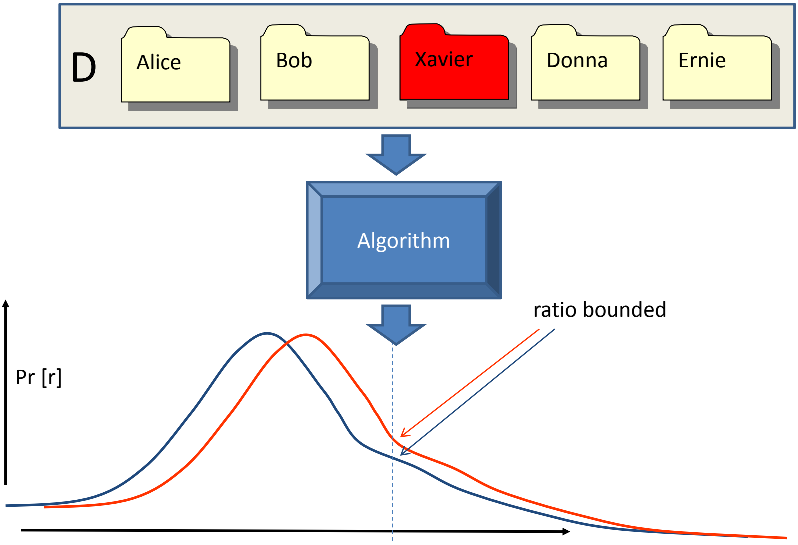
Not private

- ~~MW distributions converge to approximate equilibrium~~
- Empirical distributions also converge to approximate equilibrium

Distribution of actual plays

- Samples from MW distribution: private?

(Standard) Differential privacy [DMNS'06]



Computing equilibrium strategy privately

Idea: use distribution over plays [FS'96]

- Both players use multiplicative weights

Not private

- ~~MW distributions converge to approximate equilibrium~~
- Empirical distributions also converge to approximate equilibrium

Distribution of actual plays

- Samples from MW distribution: private?

Computing equilibrium strategy privately

Idea: use distribution over plays [FS'96]

- Both players use multiplicative weights

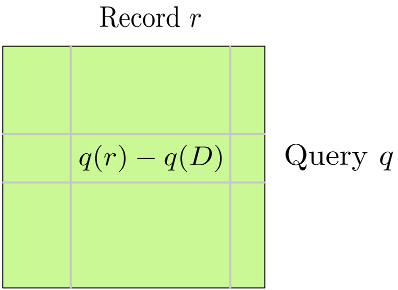
Not private

- ~~MW distributions converge to approximate equilibrium~~
- Empirical distributions also converge to approximate equilibrium

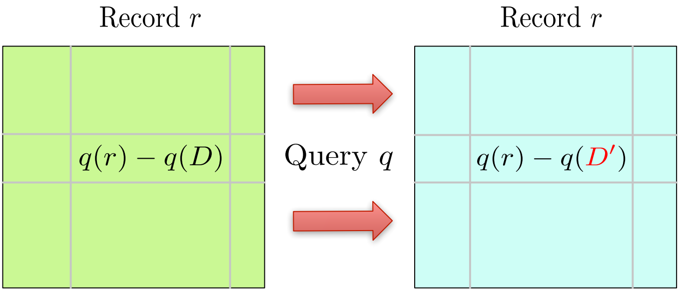
Distribution of actual plays

- Samples from MW distribution: private?
- Depends on losses: what if we change database or query?

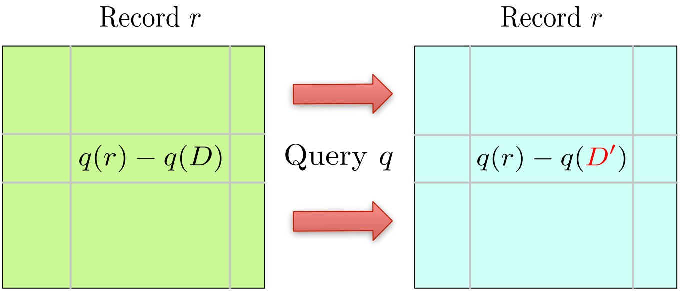
Data privacy



Data privacy

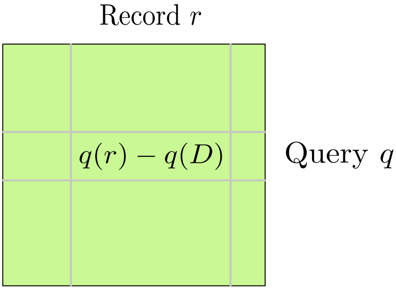


Data privacy

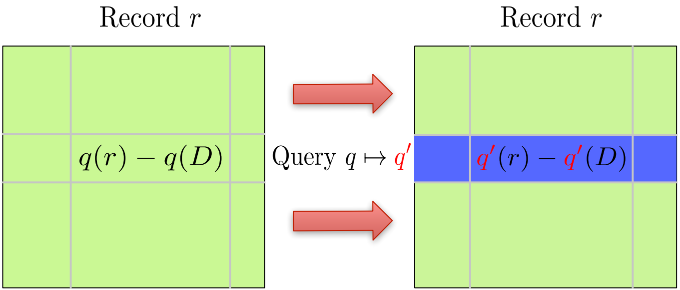


- Changing a record in database changes all losses only a little

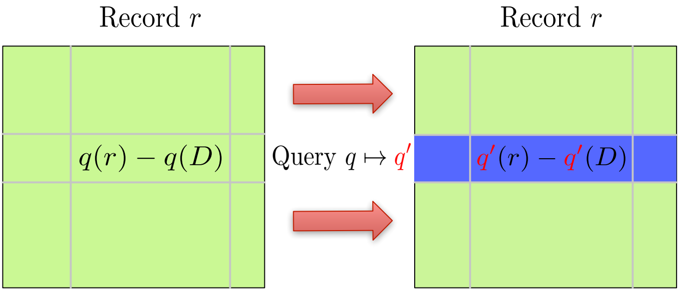
Analyst privacy



Analyst privacy



Analyst privacy



- Changing a query changes losses for an entire row (maybe by a lot)

Plan

- Private inputs: database D, set of all queries $\mathcal{Q}$ from analysts

Query release mechanism

Plan

- Private inputs: database D, set of all queries $\mathcal{Q}$ from analysts
- Simulate repeated play of query release game

Plan

- Private inputs: database D, set of all queries $\mathcal{Q}$ from analysts
- Simulate repeated play of query release game
- Publish: empirical distribution on data player's plays

Query release mechanism

Plan

- Private inputs: database D, set of all queries $\mathcal{Q}$ from analysts
- Simulate repeated play of query release game
- Publish: empirical distribution on data player's plays
- Analysts compute answers by using this as synthetic database

Requirement: Analyst privacy

- If query changed, synthetic database shouldn't change much

Requirement: Analyst privacy

- If query changed, synthetic database shouldn't change much

Obstacle: query player can't play a query too often

- Changing it might drastically change synthetic database

A closer look at the MW update

Data player's update

- Versus query q, update probability of record r:

$$p_r := p_r \cdot \exp\{-(q(r) - q(D))\}$$

Data player's update

- Versus query q, update probability of record r:

$$p_r := p_r \cdot \exp\{-(q(r) - q(D))\}$$

- After queries

$$q^{(1)}$$

$$p_r \sim \exp\Big\{ - \Big(q^{(1)}(r) - q^{(1)}(D)\Big) \Big\}$$

A closer look at the MW update

Data player's update

- Versus query q, update probability of record r:

$$p_r := p_r \cdot \exp\{-(q(r) - q(D))\}$$

- After queries

$$q^{(1)}, q^{(2)}$$

$$p_r \sim \exp\left\{ -\left(q^{(1)}(r) - q^{(1)}(D)\right) - \left(q^{(2)}(r) - q^{(2)}(D)\right) \right\}$$

A closer look at the MW update

Data player's update

- Versus query q, update probability of record r:

$$p_r := p_r \cdot \exp\{-(q(r) - q(D))\}$$

- After queries
$$q^{(1)}, q^{(2)}, \ldots, q^{(T)} :$$

$$p_r \sim \exp\Big\{ - \Big(q^{(1)}(r) - q^{(1)}(D)\Big) - \cdots - \Big(q^{(T)}(r) - q^{(T)}(D)\Big)\Big\}$$

A closer look at the MW update

Data player's update

- Versus query q, update probability of record r:

$$p_r := p_r \cdot \exp\{-(q(r) - q(D))\}$$

- After queries

$$q^{(1)}, q^{(2)}, \ldots, q^{(T)} :$$

$$p_r \sim \exp\left\{ - \left(q^{(1)}(r) - q^{(1)}(D) \right) - \cdots - \left(q^{(T)}(r) - q^{(T)}(D) \right) \right\}$$

- Very sensitive to changing a query if query played many times

Requirement: Analyst privacy

- If query changed, synthetic database shouldn't change much

Obstacle: query player can't play a query too often

- Changing it might drastically change synthetic database

Requirement: Analyst privacy

- If query changed, synthetic database shouldn't change much

Obstacle: query player can't play a query too often

- Changing it might drastically change synthetic database
- Project query distribution so probabilities are capped

Requirement: Analyst privacy

- If query changed, synthetic database shouldn't change much

Obstacle: query player can't play a query too often

- Changing it might drastically change synthetic database
- Project query distribution so probabilities are capped

No query played too often

Analyst private mechanism

Analyst private mechanism

- Maintain distributions over records and queries

Analyst private mechanism

- Maintain distributions over records and queries
- Loop:
  - Draw actions (record and query) from distributions

Putting it all together

Analyst private mechanism

- Maintain distributions over records and queries
- Loop:
  - Draw actions (record and query) from distributions
  - Calculate loss defined by the plays

Analyst private mechanism

- Maintain distributions over records and queries
- Loop:
  - Draw actions (record and query) from distributions
  - Calculate loss defined by the plays
  - Update distributions (MW)

Putting it all together

Analyst private mechanism

- Maintain distributions over records and queries
- Loop:
    - Draw actions (record and query) from distributions
    - Calculate loss defined by the plays
    - Update distributions (MW)
    - Project query distribution to cap probabilities

Putting it all together

Analyst private mechanism

- Maintain distributions over records and queries
- Loop:
  - Draw actions (record and query) from distributions
  - Calculate loss defined by the plays
  - Update distributions (MW)
  - Project query distribution to cap probabilities
- Output data's empirical distribution: synthetic database

Mishandled queries

- What if only a few queries with high error?

Changing the game

Mishandled queries

- What if only a few queries with high error?
- Query player might not be able to put high probability on these queries

Mishandled queries

- What if only a few queries with high error?
- Query player might not be able to put high probability on these queries

Probabilities are capped!

Mishandled queries

Probabilities are capped!

- What if only a few queries with high error?
- Query player might not be able to put high probability on these queries
- At equilibrium, a few queries might have high error

Analyst private mechanism

- Maintain distributions over records and queries
- Loop:
  - Draw actions (record and query) from distributions
  - Calculate loss defined by the plays
  - Update distributions (MW)
  - Project query distribution to cap probabilities
- Output data's empirical distribution: synthetic database

Putting it all together

Analyst private mechanism

- Maintain distributions over records and queries
- Loop:
  - Draw actions (record and query) from distributions
  - Calculate loss defined by the plays
  - Update distributions (MW)
  - Project query distribution to cap probabilities
- Output data's empirical distribution: synthetic database
- Find and answer queries where synthetic data performs poorly

Theorem

Suppose the analysts ask queries $\mathcal{Q}$, and let the database have n records from $\mathcal{X}$. There exists an ϵ analyst and data private mechanism which achieves error α on all queries in $\mathcal{Q}$, where

$$\alpha = O\left(\frac{\mathrm{polylog}(|\mathcal{X}|, |\mathcal{Q}|)}{\epsilon\sqrt{n}}\right).$$

Theorem

Suppose the analysts ask queries $\mathcal{Q}$, and let the database have n records from $\mathcal{X}$. There exists an ϵ analyst and data private mechanism which achieves error α on all queries in $\mathcal{Q}$, where

$$\alpha = O\left(\frac{\mathrm{polylog}(|\mathcal{X}|, |\mathcal{Q}|)}{\epsilon\sqrt{n}}\right).$$

Notes

- Counting queries, so error $\alpha \ll 1$ is nontrivial

Theorem

Suppose the analysts ask queries $\mathcal{Q}$, and let the database have n records from $\mathcal{X}$. There exists an ϵ analyst and data private mechanism which achieves error α on all queries in $\mathcal{Q}$, where

$$\alpha = O\left(\frac{\mathrm{polylog}(|\mathcal{X}|, |\mathcal{Q}|)}{\epsilon\sqrt{n}}\right).$$

Notes

- Counting queries, so error $\alpha \ll 1$ is nontrivial
- Improved dependence on n compared to $O(1/n^{1/4})$ [DNV'12], but analyst privacy guarantees are incomparable

Theorem

Suppose the analysts ask queries $\mathcal{Q}$, and let the database have n records from $\mathcal{X}$. There exists an ϵ analyst and data private mechanism which achieves error α on all queries in $\mathcal{Q}$, where

$$\alpha = O\left(\frac{\mathrm{polylog}(|\mathcal{X}|, |\mathcal{Q}|)}{\epsilon\sqrt{n}}\right).$$

Notes

- Counting queries, so error $\alpha \ll 1$ is nontrivial
- Improved dependence on n compared to $O(1/n^{1/4})$ [DNV'12], but analyst privacy guarantees are incomparable
- $O(1/\sqrt{n})$ nearly optimal dependence on n, even for data privacy only

Extensions

- One-analyst-to-many-analyst private mechanism: one analyst is allowed to change all of their queries
- Analyst private online mechanism
- Analyst private mechanism for general low-sensitivity queries

Our contributions

- Interpretation of query release as zero-sum game
- Method for privately computing the approximate equilibrium
- Nearly optimal error for one-query-to-many-analyst privacy

Wrapping up

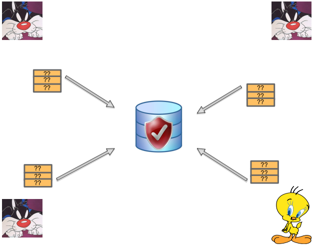
Our contributions

- Interpretation of query release as zero-sum game
- Method for privately computing the approximate equilibrium
- Nearly optimal error for one-query-to-many-analyst privacy

Ongoing/Future Work

- Inherent gap between analyst privacy and just data privacy?
- Other applications of privately solving zero-sum games?
- Solving linear programs?

Private Equilibrium Computation for Analyst Privacy



Justin Hsu, Aaron Roth,[1] Jonathan Ullman[2]

[1]University of Pennsylvania
[2]Harvard University

June 2, 2013