Differentially Private Data Cubes: Optimizing Noise Sources and Consistency

Bolin Ding (UIUC), Marianne Winslett (ADSC&UIUC), Jiawei Han (UIUC), Zhenhui Li (UIUC)

SIGMOD'11, Athens, Greece

Outline

- Introduction
  - Data cube and privacy concerns
  - Differential privacy (DP)
- Optimizing noise sources in DP publishing
- Enforcing consistency
- Experiments and future work

Outline

- Introduction
  - Data cube and privacy concerns
  - Differential privacy (DP)
- Optimizing noise sources in DP publishing
- Enforcing consistency
- Experiments and future work

Introduction

- Data cube
  - Fact table
  - Cuboids
  - Cells
  - Measure (count)

Sex	Age	Salary
F	21-30	10-50k
F	21-30	10-50k
F	31-40	50-200k
F	41-50	500k+
M	21-30	10-50k
M	21-30	50-200k
M	31-40	50-200k
M	60+	500k+

(a) Fact Table T

Sex	Age	Salary	c
*	*	0-10k	0
*	*	10-50k	3
*	*	50-200k	3
*	*	...	...

(b) Cuboid {Salary}

Sex	Age	Salary	c
F	21-30	0-10k	0
F	21-30	10-50k	2
...	...	...	...

(c) Cuboid {Sex, Age, Salary}

Sex	Age	Salary	c
*	21-30	0-10k	0
*	21-30	10-50k	3
*	21-30	50-200k	1
*	21-30	200-500k	0
*	21-30	500k+	0
*	31-40	0-10k	0
*	31-40	10-50k	0
*	31-40	50-200k	2
*	31-40	200-500k	0
*	31-40	500k+	0
*	...	...	...

(d) Cuboid {Age, Salary}

  - Application: fast OLAP, decision support, summarization

Introduction

- Privacy concerns of publishing data cube
  - Health summary tables, census data …

Sex	Age	Salary
F	21-30	10-50k
F	21-30	10-50k
F	31-40	50-200k
F	41-50	500k+
M	21-30	10-50k
M	21-30	50-200k
M	31-40	50-200k
M	60+	500k+

(a) Fact Table T

Sex	Age	Salary	c
$*$	$*$	0-10k	0
$*$	$*$	10-50k	3
$*$	$*$	50-200k	3
$*$	$*$	…	…

(b) Cuboid {Salary}

Sex	Age	Salary	c
F	21-30	0-10k	0
F	21-30	10-50k	2
…	…	…	…

(c) Cuboid {Sex, Age, Salary}

Sex	Age	Salary	c
$*$	21-30	0-10k	0
$*$	21-30	10-50k	3
$*$	21-30	50-200k	1
$*$	21-30	200-500k	0
$*$	21-30	500k+	0
$*$	31-40	0-10k	0
$*$	31-40	10-50k	0
$*$	31-40	50-200k	2
$*$	31-40	200-500k	0
$*$	31-40	500k+	0
$*$	…	…	…

(d) Cuboid {Age, Salary}

Alice with Age 31-40

Introduction

- Privacy concerns of publishing data cube
  - Health summary tables, census data …

Sex	Age	Salary
F	21-30	10-50k
F	21-30	10-50k
F	31-40	50-200k
F	41-50	500k+
M	21-30	10-50k
M	21-30	50-200k
M	31-40	50-200k
M	60+	500k+

(a) Fact Table T

Sex	Age	Salary	c
*	*	0-10k	0
*	*	10-50k	3
*	*	50-200k	3
*	*	…	…

(b) Cuboid {Salary}

Sex	Age	Salary	c
F	21-30	0-10k	0
F	21-30	10-50k	2
…	…	…	…

(c) Cuboid {Sex, Age, Salary}

Bob with Age 21-30

Sex	Age	Salary	c
*	21-30	0-10k	0
*	21-30	10-50k	3
*	21-30	50-200k	1
*	21-30	200-500k	0
*	21-30	500k+	0
*	31-40	0-10k	0
*	31-40	10-50k	0
*	31-40	50-200k	2
*	31-40	200-500k	0
*	31-40	500k+	0
*	…	…	…

75%

25%

(d) Cuboid {Age, Salary}

Introduction

- Privacy concerns of publishing data cube
  - Health summary tables, census data …
  - Adversary with sufficient background knowledge

Sex	Age	Salary
F	21-30	10-50k
F	21-30	10-50k
F	31-40	50-200k
F	41-50	500k+
M	21-30	10-50k
M	21-30	50-200k
M	31-40	50-200k
M	60+	500k+

(a) Fact Table T

Sex	Age	Salary	c
*	*	0-10k	0
*	*	10-50k	3
*	*	50-200k	3
*	*	…	…

(b) Cuboid {Salary}

Sex	Age	Salary	c
F	21-30	0-10k	0
F	21-30	10-50k	2
…	…	…	…

(c) Cuboid {Sex, Age, Salary}

Bob with Age 21-30

Sex	Age	Salary	c
*	21-30	0-10k	0
*	21-30	10-50k	3
*	21-30	50-200k	1
*	21-30	200-500k	0
*	21-30	500k+	0
*	31-40	0-10k	0
*	31-40	10-50k	0
*	31-40	50-200k	2
*	31-40	200-500k	0
*	31-40	500k+	0
*	…	…	…

(d) Cuboid {Age, Salary}

100%
0%

Carl with Age 21-30 and Salary 50-200k

Outline

- Introduction
  - Data cube and privacy concerns
  - Differential privacy (DP)
- Optimizing noise sources in DP publishing
- Enforcing consistency
- Experiments and future work

Formal Definition of DP [DworkMNS06]

An algorithm K is ϵ-differentially private if:

- for any two neighboring tables differing at most one row

$$|(T_1 - T_2) \cup (T_2 - T_1)| = 1$$

- for any set *S* of possible output

$$1 - \epsilon \approx \exp(-\epsilon) \leq \frac{\Pr\left[\mathcal{K}(T_1) \in S\right]}{\Pr\left[\mathcal{K}(T_2) \in S\right]} \leq \exp(\epsilon) \approx 1 + \epsilon$$

- Implication:
  - Any individual's record has negligible impact on query result
  - An adversary cannot make meaningful inferences about any one individual's record value

Achieving ϵ-Differential Privacy [DworkMNS06]

Query result *F:* {Tables} → R^n is a n-dim vector

Sensitivity of *F*: $S(F) = \max_{\forall \text{ neighboring } T_1, T_2} ||F(T_1) - F(T_2)||_1$

Publishing: $\tilde{F}(T) = F(T) + \langle \mathrm{Lap}(S(F)/\epsilon) \rangle^n$ is ϵ-differentially private

Density, Expectation, and Variance of Lap(l):

$$f(x) = \frac{1}{2\lambda} e^{-|x|/\lambda}$$

$$\mathrm{E}[Y] = 0$$

$$\mathrm{Var}[Y] = 2\lambda^2$$

Outline

- Introduction
  - Data cube and privacy concerns
  - Differential privacy (DP)
- Optimizing noise sources in DP publishing
- Enforcing consistency
- Experiments and future work

Approach I: Adding Noise in All Cuboids

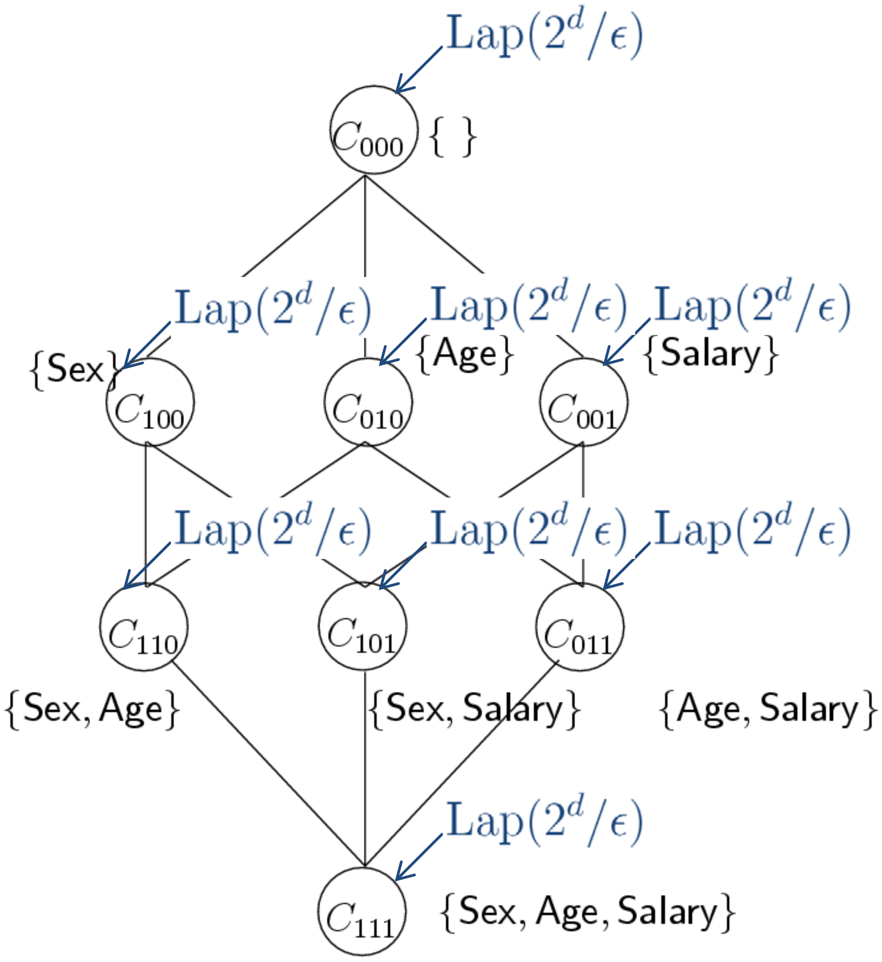
Approach All: [BarakCDKMT07]

- In a d-dim fact table
- 2^d cuboids in total
- Add noise to each of them
- Sensitivity 2^d

Max noise variance

(expected squared error):

$$\boxed{2 \times 4^d/\epsilon^2}$$

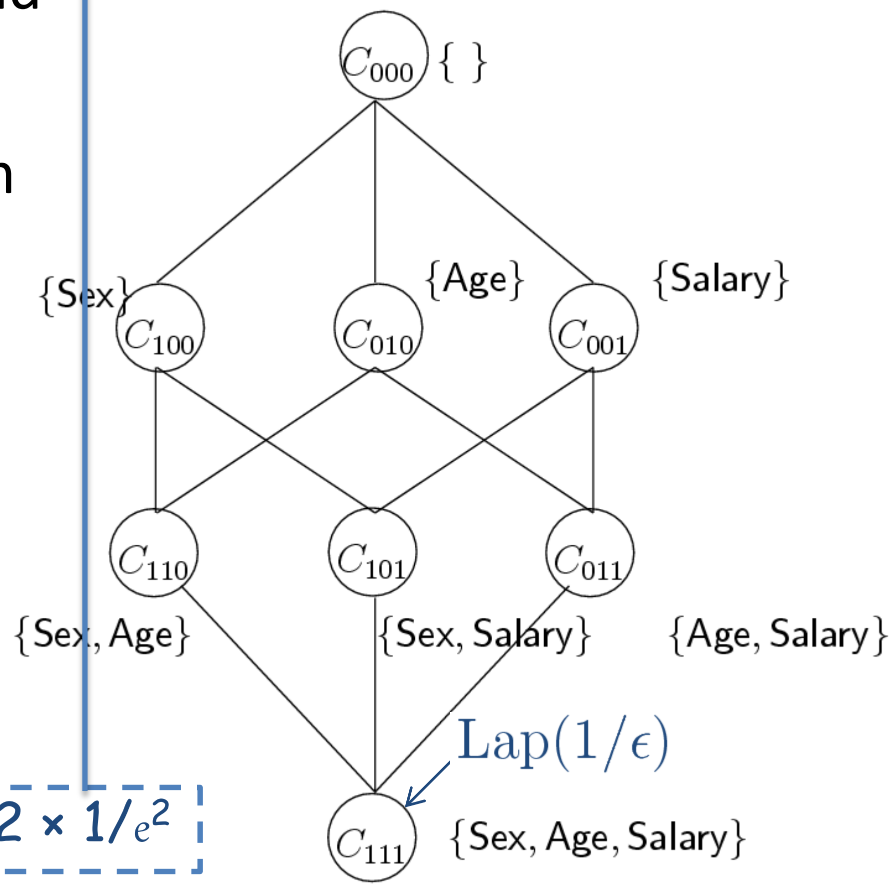


Approach II: Adding Noise in Base Cuboid

Approach Base:

- Add noise to only base cuboid
- Sensitivity 1
- Compute other cuboids from
  - the noisy base cuboid without touching the fact table (thus ϵ-DP)

$2 \times |Sex| \times |Age| \times |Salary| / \epsilon^2$



Max noise variance (expected squared error): $2 \times 1/\epsilon^2$

Can We Do Better?

Choose a set of *s* cuboids L_{pre}

Add noise to them

Sensitivity $s = |L_{pre}|$

Compute other cuboids

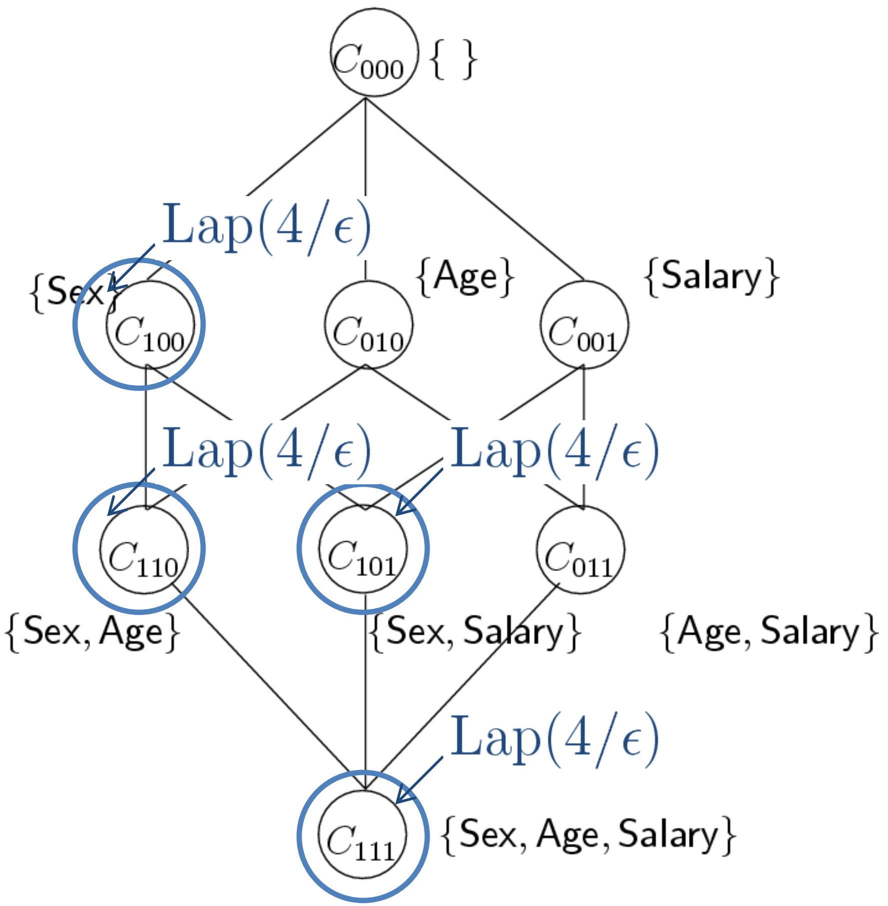
from noisy cuboids in L_{pre}

without touching the

fact table (thus ϵ-DP)

Both measure and noise

are aggregated…



Noise Aggregation

Suppose |Sex| = 2, |Age| = 7, and |Salary| = 5

Computing cuboid {Age, Salary}

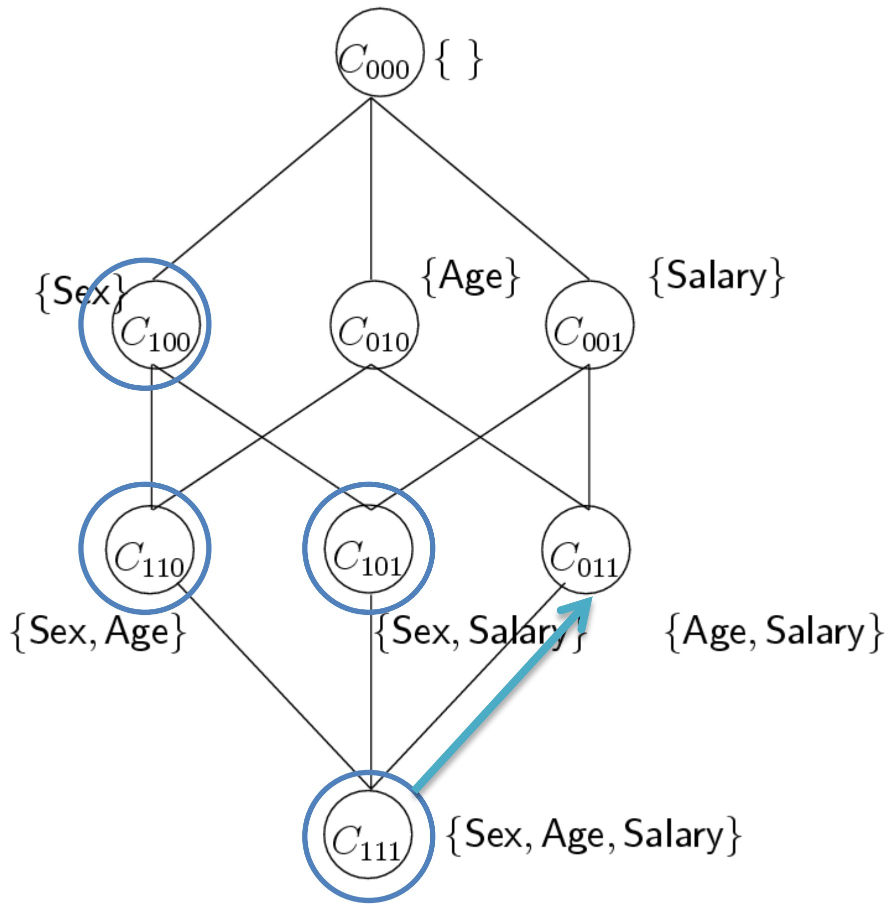
from {Sex, Age, Salary}

Sex	Age	Salary	c
F	21-30	10-50k	2
M	21-30	10-50k	1

Aggregate 2 cells in {Sex, Age, Salary}
for each cell in {Age, Salary}

Age	Salary	c
21-30	10-50k	3

Noise variance magnification: 2



Noise Aggregation

Suppose |Sex| = 2, |Age| = 7, and |Salary| = 5

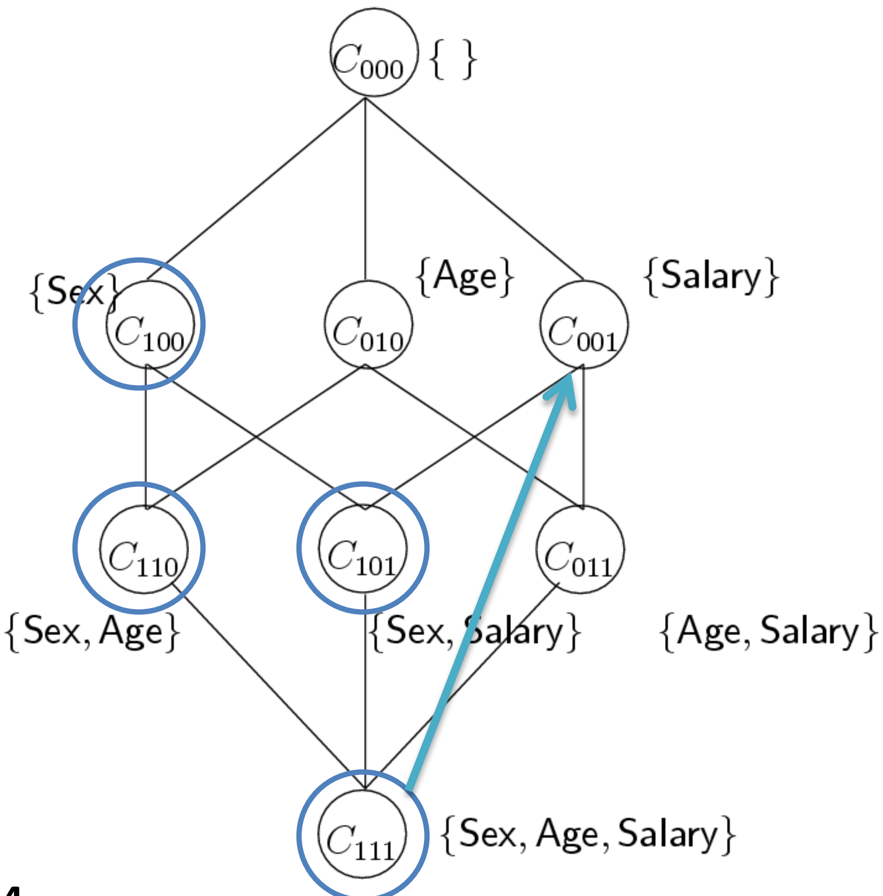
Computing cuboid {Salary}

from {Sex, Age, Salary}

Sex	Age	Salary	c
F	0-10	10-50k	0
F	…	10-50k	…
F	60+	10-50k	2
M	0-10	10-50k	0
M	…	10-50k	…
M	60+	10-50k	1

Aggregate 2×7 cells in {Sex, Age, Salary} for each cell in {Salary}

Salary	c
10-50k	12



Noise variance magnification: 14

Noise Aggregation

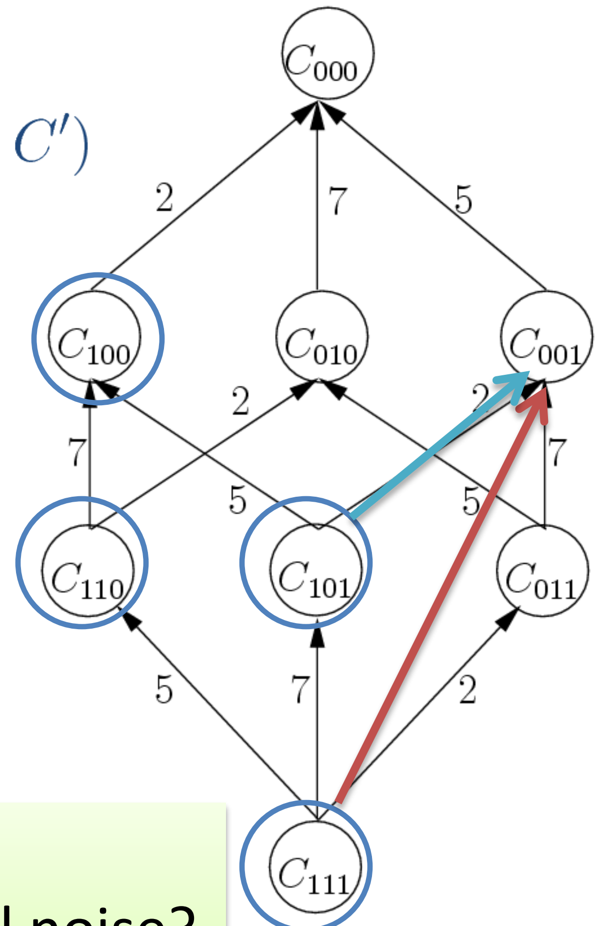
Noise variance magnification ratio of computing cuboid *C* from *C'*: $\mathrm{mag}(C, C')$

Compute *C* from a set of *s* cuboids L_{pre}

The noise variance is:

(choosing the best *C'* from L_{pre})

$$\mathrm{noise}(C, \mathcal{L}_{\text{pre}}) = \min_{C' \in \mathcal{L}_{\text{pre}}} \mathrm{mag}(C, C') \cdot 2s^2/\epsilon^2$$



A smart choice of L_{pre} can reduce the overall noise?

Optimizing Noise Sources L_{pre}

- Problem 1 (Bound Max Variance)
  - Choosing L_{pre} s.t. max noise in all cuboids
    $\mathrm{noise}(\mathcal{L}_{\mathsf{pre}}) = \max_{C} \ \mathrm{noise}(C, \mathcal{L}_{\mathsf{pre}})$ is minimized
- Problem 2 (Publish Most)
  - Given noise variance threshold q_0 and cuboid weights w
  - Choosing L_{pre} s.t. weight of precise cuboids
    $\sum_{C:\ \mathrm{noise}(C,\mathcal{L}_{\mathsf{pre}}) \le \theta_0} w(C)$ is maximized
- Problems 1 and 2 are NP-Hard
  - Reduction from **Vertex Cover In Degree-3 Graphs**
  - Design approximation algorithms

Approximation Algorithm

- Guess the optimal solution q = OPT and *s* = $|L_{pre}^{*}|$
  - using binary search
  - Fixing q and *s*

$$\mathrm{noise}(C, \mathcal{L}_{\mathsf{pre}}) \le \theta \Leftrightarrow \min_{C' \in \mathcal{L}_{\mathsf{pre}}} \mathrm{mag}(C, C') \le \frac{\theta\epsilon^2}{2s^2}$$

  - Define coverage of a cuboid *C'*

$$\mathrm{cov}(C') = \{C \in \mathcal{L} \mid C \preceq C',\ \mathrm{mag}(C, C') \le \frac{\theta\epsilon^2}{2s^2}\}$$

  - Sub-problem: Select *s* cuboids L_{pre} to cover all cuboids L

Approximation Algorithm

- Guess q and s
- Solve sub-problem:
  - Select s cuboids L_{pre} to cover all cuboids L

  Using the greedy algorithm for Set Cover
  - We may need $(\log|L|+1)s$ cuboids
  - So noise is magnified another $(\log|L|+1)^2$ times
- So, $(\log|L|+1)^2$-approximation

Outline

- Introduction
  - Data cube and privacy concerns
  - Differential privacy (DP)
- Optimizing noise sources in DP publishing
- Enforcing consistency
- Experiments and future work

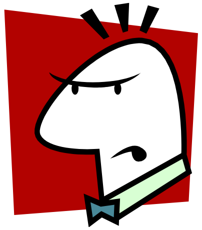
Enforcing Consistency

- Possible inconsistency
  - Independent noise

Sex	Age	Salary	c
F	21-30	10-50k	2 + 0.5 = 2.5
M	21-30	10-50k	1 – 0.2 = 0.8

Age	Salary	c
21-30	10-50k	3 – 0.2 = 2.8

  - A sign of bad data?



Consistency Constraints

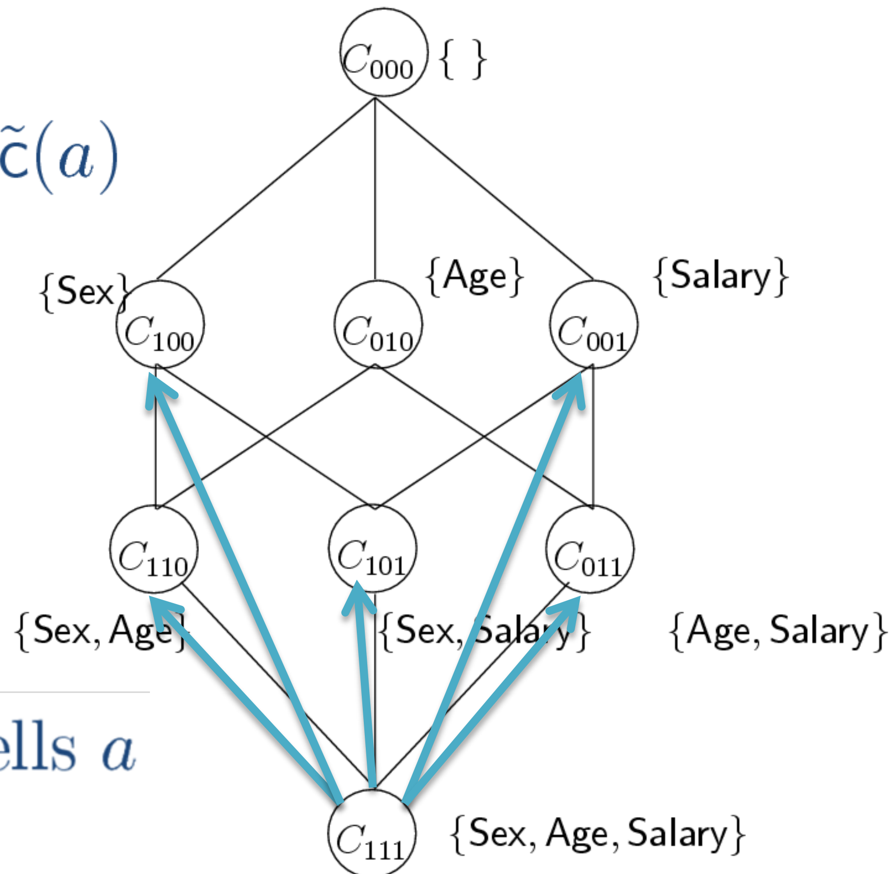
Every cuboid has the measure as if it is computed from the base cuboid

Noisy measure (ensuring DP): $\tilde{c}(a)$

Consistent measure: $\hat{c}(a)$

Subject to consistency constraints:

$$\sum_{a' \in \mathrm{Base}(a)} \hat{c}(a') = \hat{c}(a), \quad \forall \text{ cells } a$$



Consistency Constraints

Base cells under a cell a:

Sex	Age	Salary	c
F	21-30	10-50k	2
M	21-30	10-50k	1

$\text{Base}(a)$

Age	Salary	c
21-30	10-50k	3

Sex	Age	Salary	c
F	0-10	10-50k	0
F	…	10-50k	…
F	60+	10-50k	2
M	0-10	10-50k	0
M	…	10-50k	…
M	60+	10-50k	1

$\text{Base}(a)$

Salary	c
10-50k	12

Consistency-Enforcing Framework

Minimizing L^p distance between $\tilde{c}(a)$ and $\hat{c}(a)$

$$||\hat{c}(\cdot) - \tilde{c}(\cdot)||_p = \sum_{a \in \mathcal{E}_{\text{pre}}} \left(|\hat{c}(a) - \tilde{c}(a)|^p\right)^{1/p}$$

subject to consistency constraints

$\mathcal{E}_{\text{pre}}$: all cells in $\mathcal{L}_{\text{pre}}$

Intuition:
We do not know the real measure values…
Then let's approximate the noisy version

L^∞ Version

- Minimizing L^∞ distance

$$\begin{aligned} &\text{minimize } z \\ &\text{s.t.} \quad |\hat{\mathsf{c}}(a) - \tilde{\mathsf{c}}(a)| \le z, \quad \forall \text{ cells } a \in \mathcal{E}_{\mathsf{pre}}; \\ &\qquad \sum_{a' \in \mathrm{Base}(a)} \hat{\mathsf{c}}(a') = \hat{\mathsf{c}}(a), \quad \forall \text{ cells } a \in \mathcal{E}_{\mathsf{pre}}. \end{aligned}$$

Generalizing [BarakCDKMT07]

With probability at least $1-\delta$, where $\delta = \frac{|\mathcal{E}_{\mathsf{pre}}|}{e^{\eta/2}}$,

$$\sum_{a \in \mathcal{E}_{\mathsf{pre}}} |\hat{\mathsf{c}}(a) - \mathsf{c}(a)| \le \frac{|\mathcal{E}_{\mathsf{pre}}||\mathcal{L}_{\mathsf{pre}}|}{\epsilon} 2 \log \frac{|\mathcal{E}_{\mathsf{pre}}|}{\delta} = \frac{|\mathcal{E}_{\mathsf{pre}}||\mathcal{L}_{\mathsf{pre}}|}{\epsilon} \eta.$$

L^1 Version

- Minimizing L^1 distance

$$\text{minimize} \sum_{a \in \mathcal{E}_{\mathsf{pre}}} z_a$$

$$\text{s.t.} \quad |\hat{\mathsf{c}}(a) - \tilde{\mathsf{c}}(a)| \le z_a, \quad \forall \text{ cell } a \in \mathcal{E}_{\mathsf{pre}};$$

$$\sum_{a' \in \mathrm{Base}(a)} \hat{\mathsf{c}}(a') = \hat{\mathsf{c}}(a), \quad \forall \text{ cell } a \in \mathcal{E}_{\mathsf{pre}}.$$

With probability at least $1 - \delta$, where $\delta = \left(\frac{\eta}{2e^{\eta/2-1}}\right)^{|\mathcal{E}_{\mathsf{pre}}|}$,

$$\sum_{a \in \mathcal{E}_{\mathsf{pre}}} |\hat{\mathsf{c}}(a) - \mathsf{c}(a)| \le \frac{|\mathcal{E}_{\mathsf{pre}}||\mathcal{L}_{\mathsf{pre}}|}{\epsilon}\eta.$$

L^1 Version

- Analysis

$$\sum_{a\in\mathcal{E}_{\text{pre}}} |\hat{\mathsf{c}}(a)-\mathsf{c}(a)| \le \sum_{a\in\mathcal{E}_{\text{pre}}} |\hat{\mathsf{c}}(a)-\tilde{\mathsf{c}}(a)| + \sum_{a\in\mathcal{E}_{\text{pre}}} |\tilde{\mathsf{c}}(a)-\mathsf{c}(a)|$$
$$\le \sum_{a\in\mathcal{E}_{\text{pre}}} |\mathsf{c}(a)-\tilde{\mathsf{c}}(a)| + \sum_{a\in\mathcal{E}_{\text{pre}}} |\tilde{\mathsf{c}}(a)-\mathsf{c}(a)|$$
$$= 2\sum_{a\in\mathcal{E}_{\text{pre}}} |\mathsf{c}(a)-\tilde{\mathsf{c}}(a)|$$

$$\mathsf{c}(a)-\tilde{\mathsf{c}}(a) \sim \mathrm{Lap}(|\mathcal{L}_{\text{pre}}|/\epsilon) \Rightarrow |\mathsf{c}(a)-\tilde{\mathsf{c}}(a)| \sim \mathrm{Exponential}(\epsilon/|\mathcal{L}_{\text{pre}}|)$$

Extending Chernoff's Inequality to Exponential Distribution

With probability at least $1-\delta$, where $\delta = (\frac{\eta}{e^{\eta-1}})^{|\mathcal{E}_{\text{pre}}|}$,

$$\sum_{a\in\mathcal{E}_{\text{pre}}} |\mathsf{c}(a)-\tilde{\mathsf{c}}(a)| \le \mathrm{E}\left[\sum_{a\in\mathcal{E}_{\text{pre}}} |\mathsf{c}(a)-\tilde{\mathsf{c}}(a)|\right]\eta = \frac{|\mathcal{E}_{\text{pre}}||\mathcal{L}_{\text{pre}}|}{\epsilon}\eta$$

Let $X = X_1 + X_2 + \ldots + X_n$ where X_i's are i.i.d. with the exponential distribution. With probability at least $1-\delta$, where $\delta = \left(\frac{\eta}{e^{\eta-1}}\right)^n$, we have $X \le \eta\mathrm{E}[X]$ $(\eta > 1)$.

L^2 Version

Inspired by DP Range Query in [HayRMS10]

- Minimizing L^2 distance

$$\text{minimize} \sum_{a \in \mathcal{E}_{\text{pre}}} \left(\hat{\mathsf{c}}(a) - \tilde{\mathsf{c}}(a)\right)^2$$

$$\text{s.t.} \sum_{a' \in \text{Base}(a)} \hat{\mathsf{c}}(a') = \hat{\mathsf{c}}(a), \quad \forall \text{ cell } a \in \mathcal{L}_{\text{pre}}.$$

- **Surprisingly, solvable it in linear time!**

  LP is not practical in this context… Faster than OLS…

- **Statistics optimality**
  - A **unbiased** estimator of the real values of measure
  - The **smallest variance** (expected squared error) among any linear unbiased estimator

Comparing L$^\infty$, L^1, and L^2 Versions

Generalizing [BarakCDKMT07]

L$^\infty$: With probability at least $1-\delta$, where $\delta = \frac{|\mathcal{E}_{\mathsf{pre}}|}{e^{\eta/2}}$,

L^1: With probability at least $1-\delta$, where $\delta = (\frac{\eta}{2e^{\eta/2-1}})^{|\mathcal{E}_{\mathsf{pre}}|}$,

$$\sum_{a \in \mathcal{E}_{\mathsf{pre}}} |\hat{\mathsf{c}}(a) - \mathsf{c}(a)| \leq \frac{|\mathcal{E}_{\mathsf{pre}}||\mathcal{L}_{\mathsf{pre}}|}{\epsilon}\eta.$$

L^2: The smallest variance among all linear unbiased estimators
Efficient in practice (linear-time solvable)

Outline

- Introduction
  - Data cube and privacy concerns
  - Differential privacy (DP)
- Optimizing noise sources in DP publishing
- Enforcing consistency
- Experiments and future work

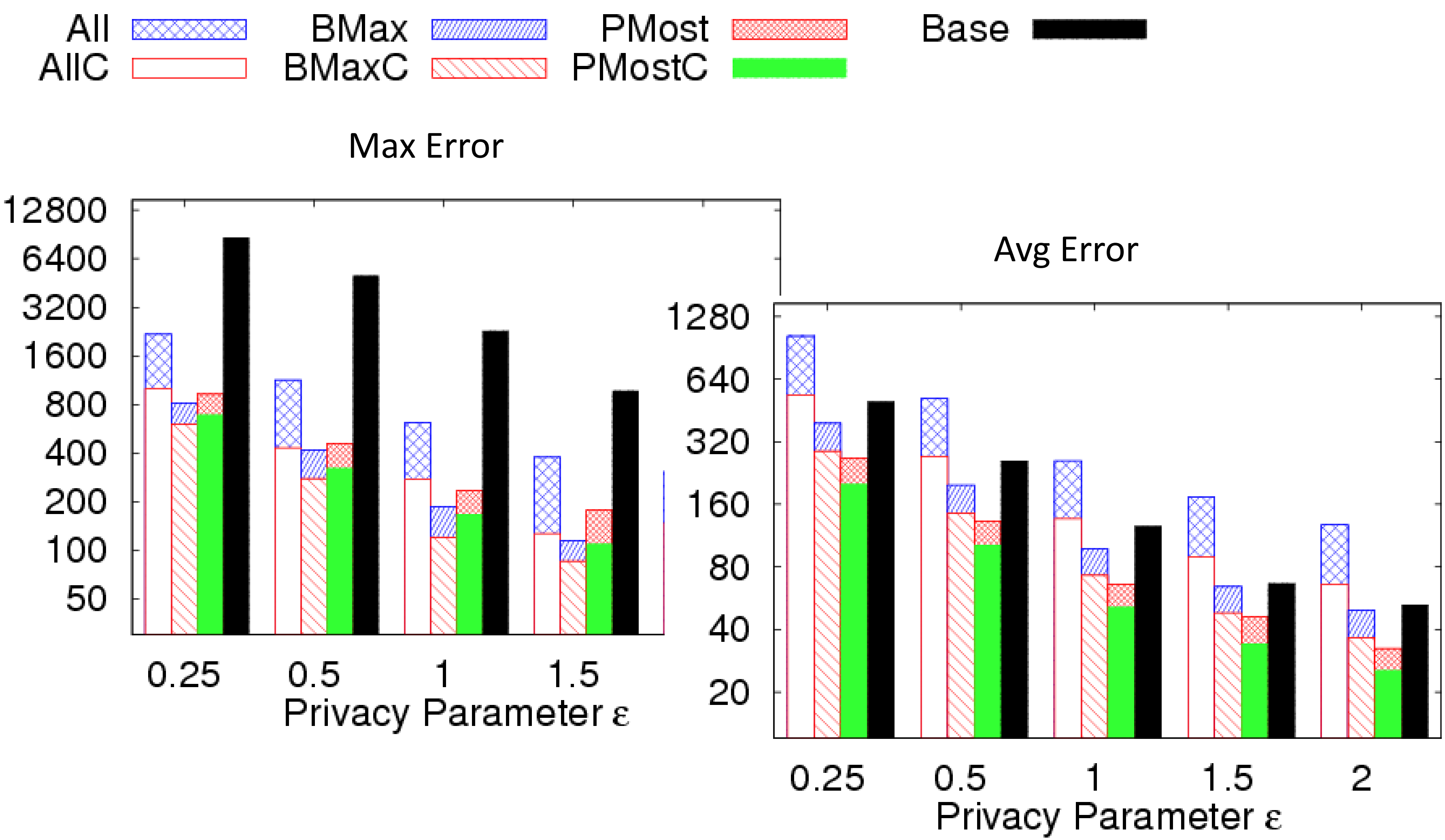
Experiments

- Seven algorithms
  - Baselines: All, Base
  - Optimizing noise sources: BMax, PMost
  - Enforcing consistency: AllC, BMaxC, PMostC

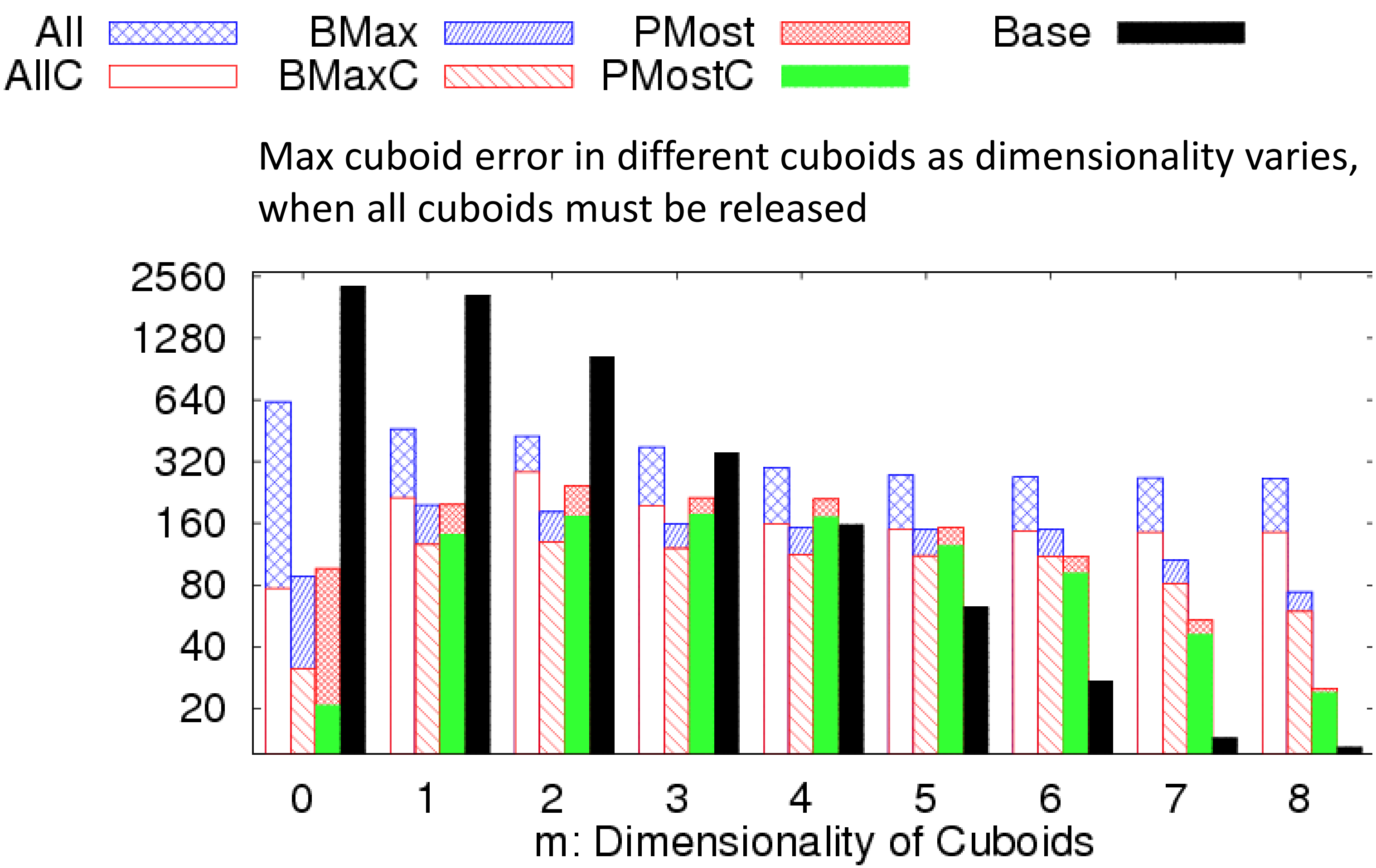
- Dataset
  - Adult dataset from http://archive.ics.uci.edu/ml/
  - 8 categorical dimensions:

    workclass (cardinality 9), education (16), marital-status (7), occupation (15), relationship (6), race (5), sex (2), and salary (2).

Experiments

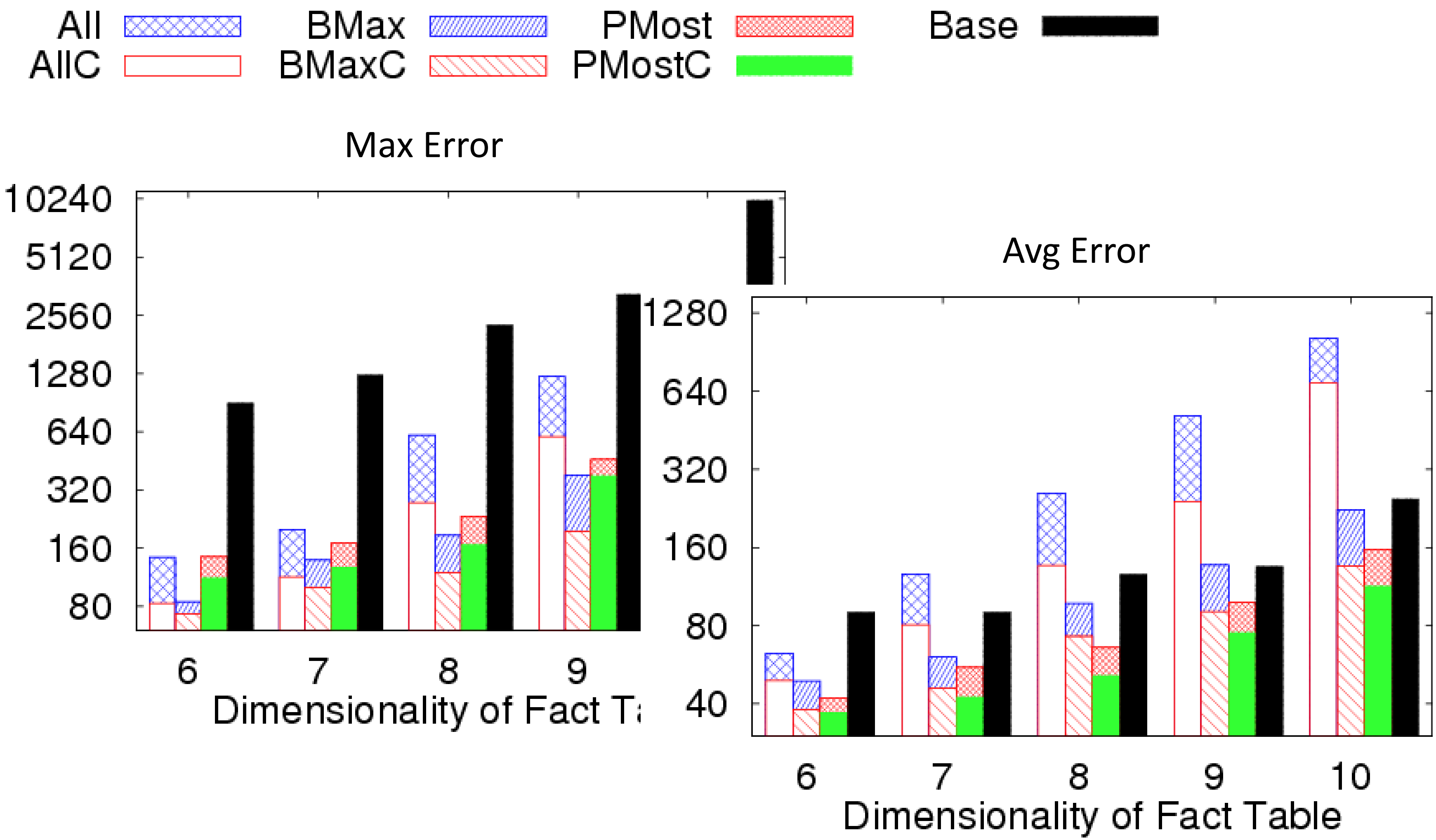


Experiments



Max cuboid error in different cuboids as dimensionality varies, when all cuboids must be released

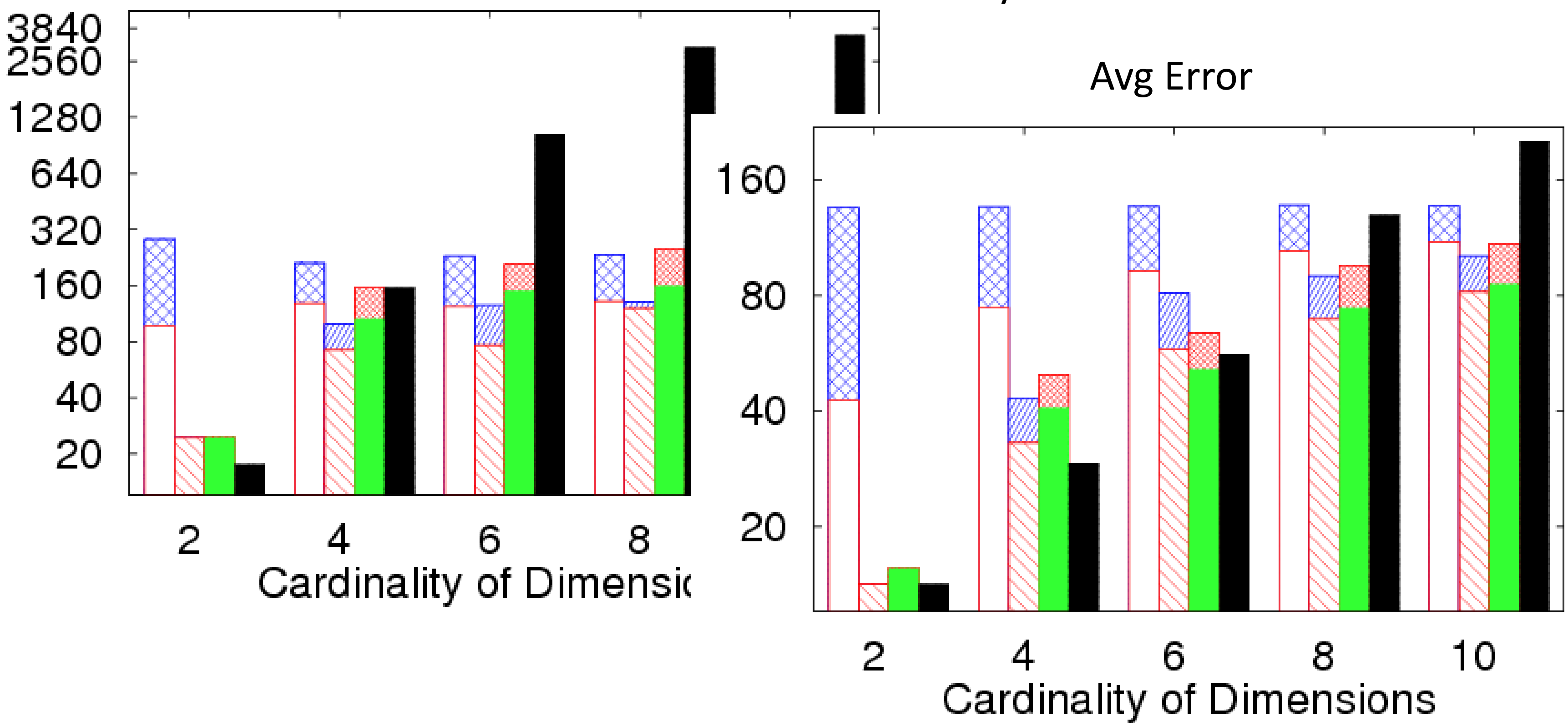
Experiments



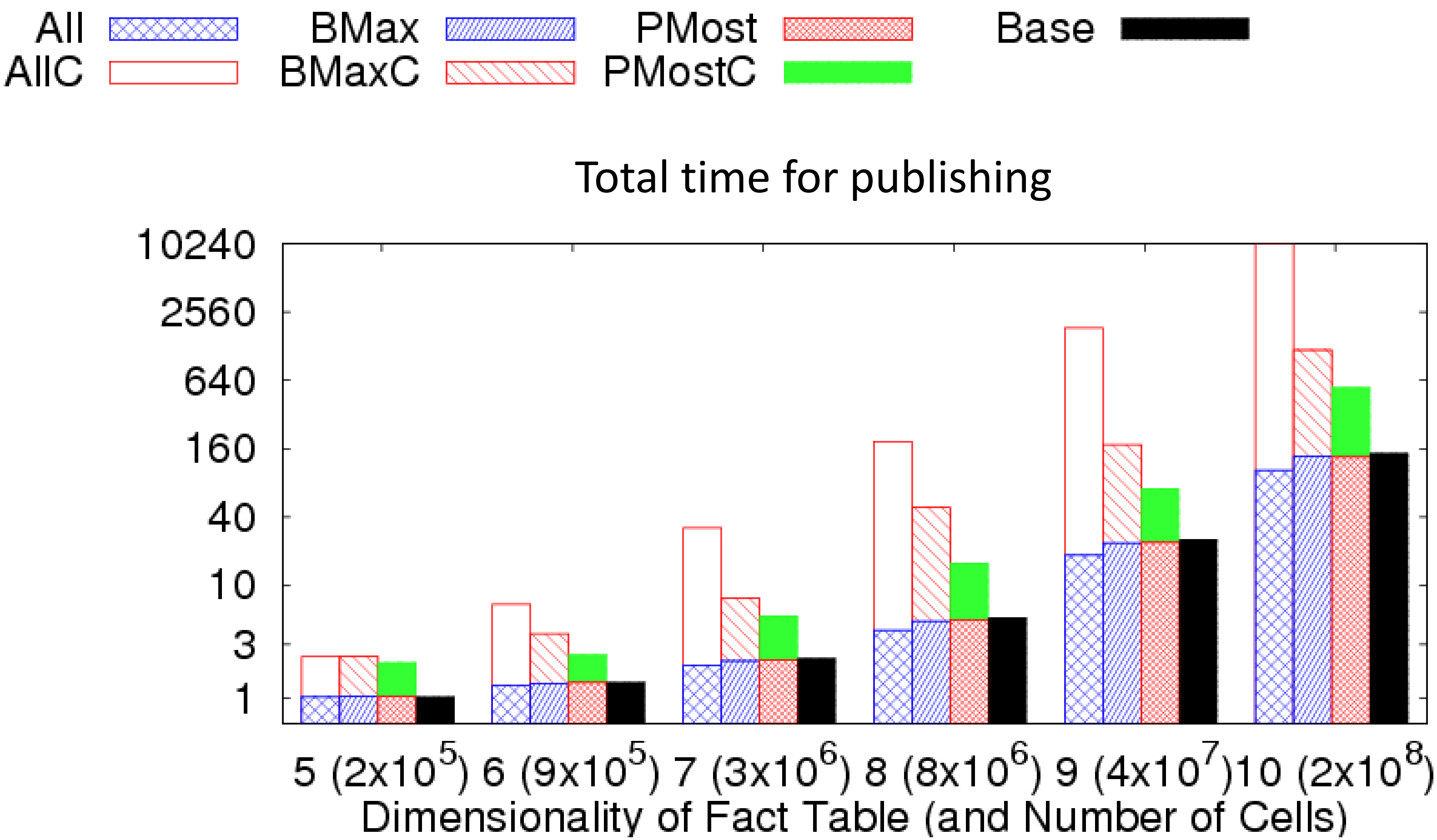
Experiments

Fix the number of dimensions (7), vary their cardinalities

Max Error

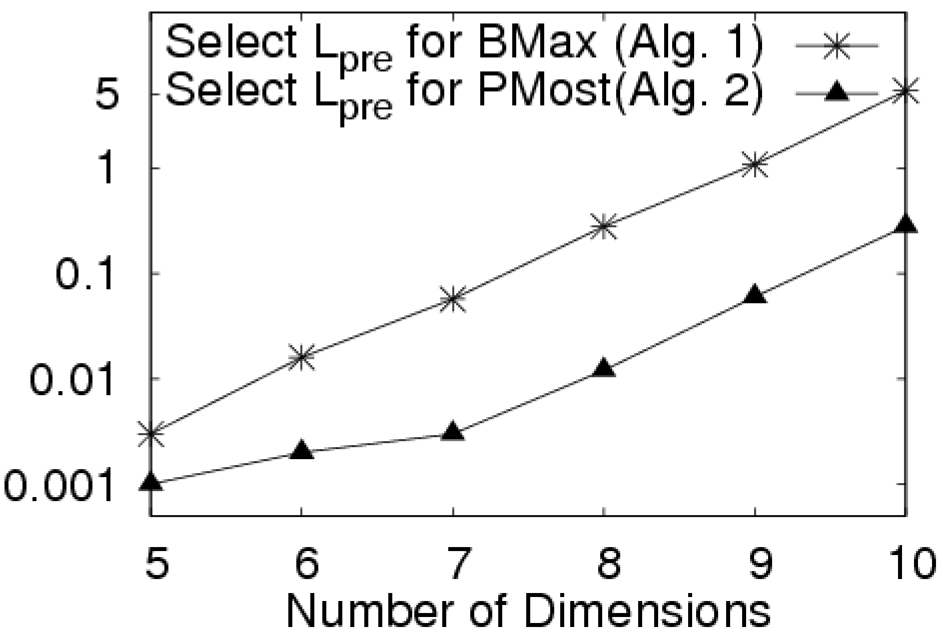


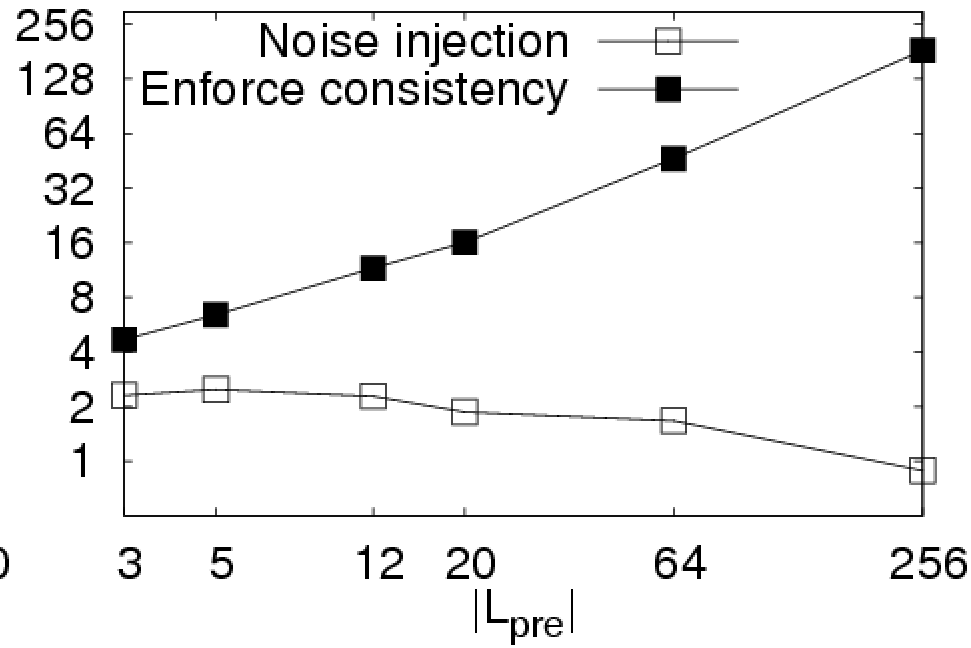
Experiments



Experiments

Running time for each subroutine





Conclusion and Future Work

- Conclusion
  - Publishing a data cube in a differentially private way
  - Optimizing noise sources in DP data publishing algorithms
  - Enforcing consistency in data cubes

- Ongoing work and open questions
  - Gap between hardness and approximation
    (better approximation algorithm?)
  - Online query model
  - Handling different classes of data cube measures
  - Some cuboids are exact while some are noisy?

Thank You!