CS208: Applied Privacy for Data Science
Implementing Differential Privacy: One-Shot Release

James Honaker & Salil Vadhan

School of Engineering & Applied Sciences
Harvard University

March 4, 2019

Beyond Noise Addition

- Recall: median over $\mathcal{X} = [0,1]$ has global sensitivity is 1.
  - Laplace Mechanism is useless.

- But on many natural datasets, the local sensitivity is small:

$$\mathrm{LS}_q(x) \stackrel{\text{def}}{=} \max_{x' : x' \sim x} |q(x') - q(x)|.$$

  - Adding noise proportional to local sensitivity is not DP.
  - But there are several DP methods that approximate this idea (smooth sensitivity, propose-test-release, privately bounding local sensitivity, restricted sensitivity).

- Laplace mechanism is only optimizing worst-case, additive accuracy for a single, real-valued query.

Exponential Mechanism

- For median:

  $M(x)$: output $y \in \mathcal{X}$ with probability $\propto \exp(\varepsilon \cdot u(x,y)/2)$,

  – Where $u(x,y) = \min\{\#\{i : x_i \le y\}, \#\{i : x_i \ge y\}\}$.

- In general:

  $M(x)$: output $y \in \mathcal{Y}$ with probability $\propto \exp\left(\varepsilon \cdot \frac{u(x,y)}{2 \cdot \mathrm{GS}_u}\right)$,

  where $\mathrm{GS}_u \stackrel{\text{def}}{=} \max_{x \sim x', y} |u(x,y) - u(x',y)|$.

Approximate Differential Privacy

Def: M is (ε, δ)-DP if for all D∼D', and all q

$\forall$ sets T, $\Pr[M(D,q) \in T] \leq e^{\varepsilon} \cdot \Pr[M(D',q) \in T] + \delta$

- Intuitively: ε-DP with probability at least $1 - \delta$.
- Picking a random person from dataset and publishing their data is $(0, 1/n)$-DP, so want $\delta \ll 1/n$.
- Ideally set δ to be cryptographically small (e.g. 2^{-50}).
- Satisfies postprocessing, basic composition (adding δ_i's).
- Group privacy for groups of size up to $O(1/\varepsilon)$.
- Does not suffice to check pointwise (need to consider sets T).

Benefits of Approximate DP

- More mechanisms, e.g. Gaussian Mechanism:

$$M(x, q) = q(x) + \mathcal{N}\left(0, \sigma^2\right),$$
$$\text{for } \sigma = \frac{\mathrm{GS}_q}{\varepsilon} \cdot \sqrt{2 \ln(2/\delta)}$$

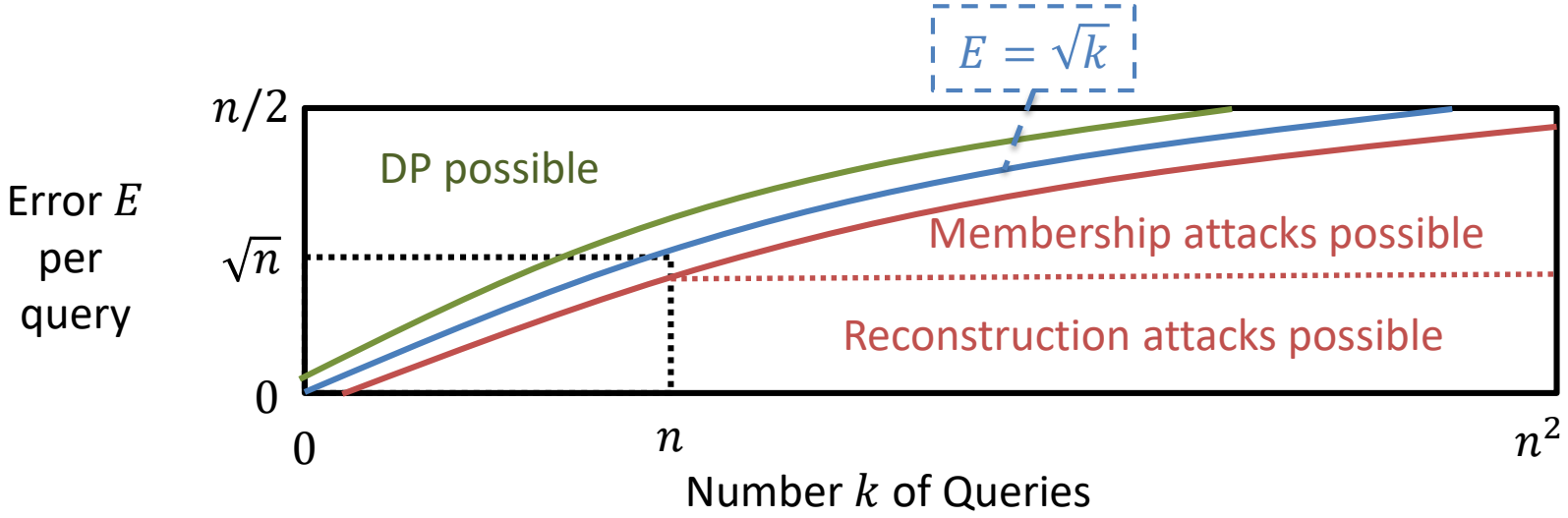
- Advanced Composition Thm: If M_i is (ε, δ)-DP for $i = 1, \ldots, k$ and $k < 1/\varepsilon^2$, then $\forall \delta > 0$

$$M\left(x, (q_1, \ldots, q_k)\right) = \left(M_1(x, q_1), \ldots, M_k(x, q_k)\right)$$
$$\text{is } (\varepsilon', k \cdot \delta + \delta') \text{ −DP, for}$$
$$\varepsilon' = O\left(\varepsilon \cdot \sqrt{k \cdot \log(1/\delta')}\right).$$

Queries vs. Accuracy Tradeoff

Using Laplace Mechanism to answer k queries, each with global sensitivity 1 (e.g. counts), under fixed privacy budget ε':

- Set $\varepsilon = 1/\tilde{O}\left(\sqrt{k}\right)$ for each query (via Advanced Comp, hiding δ').
- Add noise of scale $E = 1/\varepsilon \approx \tilde{O}\left(\sqrt{k}\right)$ per query.



Note: DP prevents all membership & reconstruction attacks (not just those we've seen), e.g. $\Pr[\text{true pos}] \leq e^{\varepsilon} \cdot \Pr[\text{false pos}] + \delta$

Doing Better than Composition

- Not all sequences of k queries require error growing as $\sqrt{k}$.

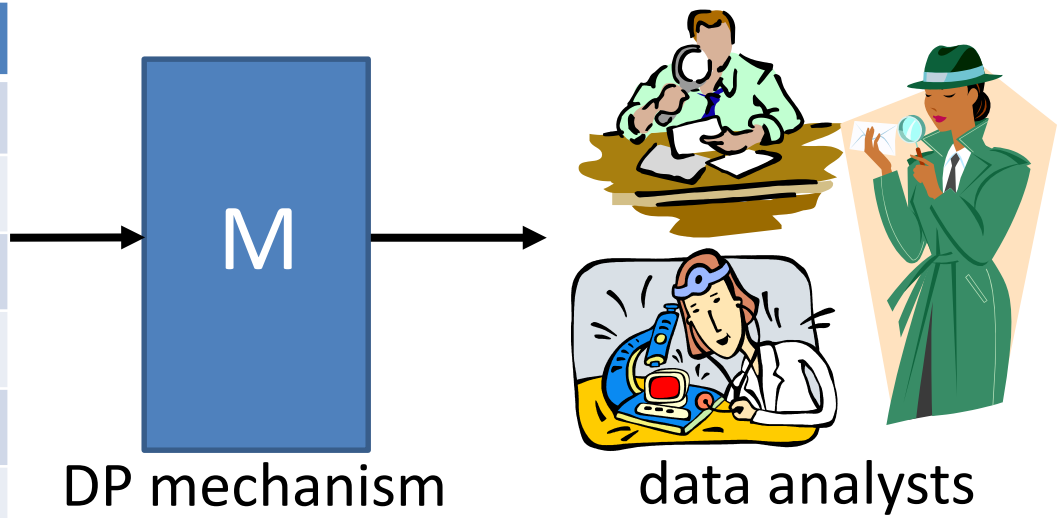
- Example: histograms
  - Let $B_1, \ldots, B_k \subseteq \mathcal{X}$ be disjoint bins.
  - Define $q_j : \mathcal{X}^n \to \{0,1\}$ by $q_j(x) = \#\{i : x_i \in B_j\}$.
  - Define $M(x) = (q_1(x) + Z_1, q_2(x) + Z_2, \ldots, q_k(x) + Z_k)$ where the Z_j's are independent $\mathrm{Lap}(2/\varepsilon)$ or $\mathrm{Geo}(2/\varepsilon)$.
  - Then M is ε-DP.

- Amazing result: with correlated noise, can answer k arbitrary bounded averaging queries on a finite data universe $\mathcal{X}$ w/error

$$\alpha = O\left(\frac{\sqrt{\log|\mathcal{X}| \cdot \log(1/\delta)} \cdot \log k}{\varepsilon n}\right)^{1/2}$$

One-Shot Releases

Sex	Blood	…	HIV?
F	B	…	Y
M	A	…	N
M	O	…	N
M	O	…	Y
F	A	…	N
M	B	…	Y

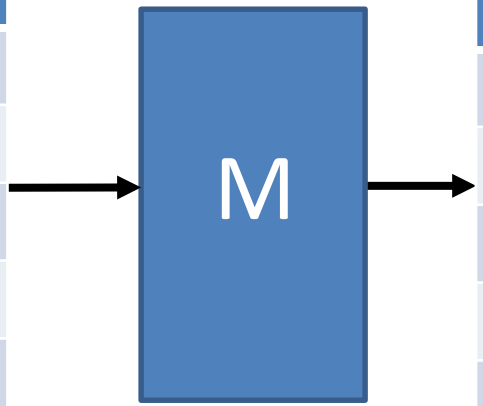


Goal: release as much useful info as possible given privacy budget

- Ideally support unforeseen analyses
- Summary statistics
- ML model
- Synthetic data

Differentially Private Synthetic Data

Sex	Blood	⋯	HIV?
F	B	⋯	Y
M	A	⋯	N
M	O	⋯	N
M	O	⋯	Y
F	A	⋯	N
M	B	⋯	Y



Sex	Blood	⋯	HIV?
M	B	⋯	N
F	B	⋯	Y
M	O	⋯	Y
F	A	⋯	N
F	O	⋯	N

M: $\mathcal{X}^n \rightarrow \mathcal{X}^m$ such that:

- $M(x)$ has the same syntax as a real dataset.
- $M(x)$ reflects many statistical properties of x.
- M is differentially private.

Synthetic Data via DP Histograms

- Use singleton bins $B_y = \{y\}$ for each $y \in \mathcal{Y}$.
- Construct a DP histogram $(a_1, \ldots, a_{|\mathcal{X}|}) \leftarrow M_{\text{hist}}(x)$, where $a_y \approx \#\{i : x_i = y\}$.
- Output synthetic dataset $\hat{x}$ with a_y copies of each element y.

Difficulties?

- a_y's may not be nonnegative integers.
  - Soln 1: use Geometric Mechanism and clamp at 0.
  - Soln 2: use Exponential Mechanism with range $\{0, \ldots, n\}$.
- Poor utility & efficiency when $\mathcal{X}$ is large.

Stability-Based Histogram

1. Let $B_1, \ldots, B_k \subseteq \mathcal{X}$ be disjoint bins.
2. Define $q_j : \mathcal{X}^n \to \{0,1\}$ by $q_j(x) = \#\{i : x_i \in B_j\}$.
3. For each j s.t. $q_j(x) > 0$:
   a) Let $a_j = q_j(x) + Z_j$ for $Z_j \sim \mathrm{Geo}(2/\varepsilon)$.
   b) If $a_j > \left\lceil \frac{2}{\varepsilon} \cdot \ln \frac{1}{\delta} \right\rceil$, output (j, a_j).
4. Treat all other bins as having a zero count.

Intuition for (ε, δ)-DP:

- Only difference from pure DP is treatment of zero bins.
- If $q_j(x) = 0$, then $q_j(x') \leq 1$ for any $x' \sim x$, and

$$\Pr\left[1 + Z_j > \left\lceil \frac{2}{\varepsilon} \cdot \ln \frac{1}{\delta} \right\rceil\right] < \delta.$$

Stability-Based Histogram

1. Let $B_1, \ldots, B_k \subseteq \mathcal{X}$ be disjoint bins.
2. Define $q_j : \mathcal{X}^n \to \{0,1\}$ by $q_j(x) = \#\{i : x_i \in B_j\}$.
3. For each j s.t. $q_j(x) > 0$:
   a) Let $a_j = q_j(x) + Z_j$ for $Z_j \sim \mathrm{Geo}(2/\varepsilon)$.
   b) If $a_j > \left\lceil \frac{2}{\varepsilon} \cdot \ln \frac{1}{\delta} \right\rceil$, output (j, a_j).
4. Treat all other bins as having a zero count.

Benefits:

- Computation and output size linear in n rather than $|\mathcal{X}|$.
- Max error $O((1/\varepsilon) \cdot \ln(1/\delta))$ whp, independent of $|\mathcal{X}|$.
- But still can have poor utility when $|\mathcal{X}|$ large. (Why?)

Census Bureau's Use of DP

Excerpts from:

- Simson Garfinkel "Challenges and Experiences Adapting Differentially Private Mechanisms to the 2020 Census," FCSM 2018.

See also:

- John Abowd. "The U.S. Census Bureau Adopts Differential Privacy," KDD 2018.
  [how to decide on privacy vs. accuracy]
- Dan Kifer. "Consistency with External Knowledge: The Top-Down Algorithm," Simons Privacy workshop TODAY.
  [many algorithmic issues and choices]
- Aref Danjani. "The modernization of statistical disclosure limitation at the U.S. Census Bureau," UNECE/EUROSTAT 2017.
  [challenges for other Census products]

Consistency & Optimization

- Structural Zeroes: Enforced by edit and imputation, DP can't reintroduce it
  - Householder and spouse/partner must be at least 15 yrsol
  - Every household must have exactly one householder
  - At least one of the binary race flags must be 1
  - Etc.

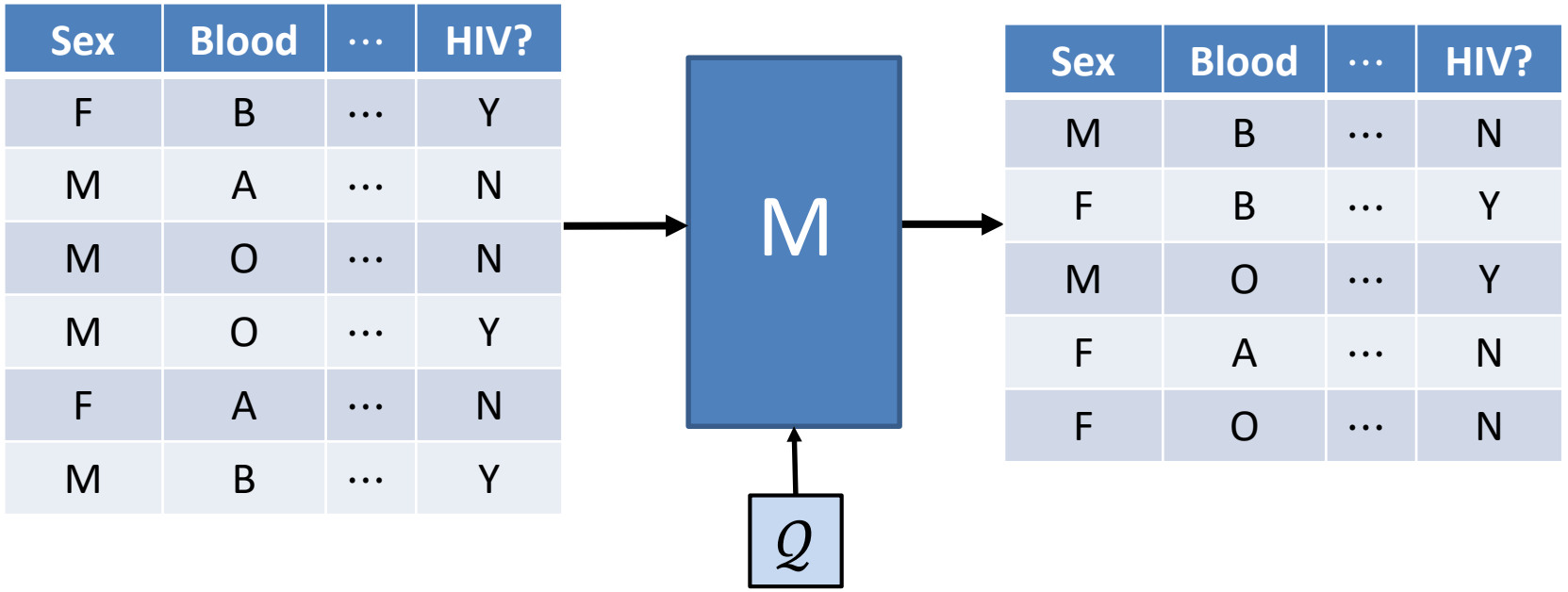
- Invariants: public statistics with exact values
  - State population totals
  - Linear constraints: sum of county populations equals state population
  - Single-gender group quarters (dorms, prisons)

- Optimizing accuracy: for a set Q of queries
  - Use "matrix mechanism" to determine related set Q' of queries, apply Laplace mechanism to Q' , then reconstruct synthetic data.
  - With constraints, NP-hard: use integer programming heuristics.

[LeClerc-Clark-Sexton `17, Kifer `19]

Private Multiplicative Weights

[Blum-Ligett-Roth `08,…,Hardt-Rothblum `10]



(ε, δ)-DP M: $\mathcal{X}^n \to \mathcal{X}^m$ such that $\forall q \in \mathcal{Q},\ q: \mathcal{X} \to [0,1]$

$$\left|\frac{1}{n}\sum_{i=1}^{n} q(x_i) - \frac{1}{m}\sum_{i=1}^{m} q(M(x)_i)\right| \leq O\left(\frac{\sqrt{\log|\mathcal{X}|\cdot\log(1/\delta)}\cdot\log|\mathcal{Q}|}{\varepsilon n}\right)^{1/2}$$

Private Mult. Weights

[Hardt-Rothblum `10]

(ε, δ)-DP $M\colon\ \mathcal{X}^n \to \mathcal{X}^m$ such that $\forall q \in \mathcal{Q},\ q\colon \mathcal{X} \to [0,1]$

$$\left|\frac{1}{n}\sum_{i=1}^{n} q(x_i) - \frac{1}{m}\sum_{i=1}^{m} q(M(x)_i)\right| \leq O\left(\frac{\sqrt{\log|\mathcal{X}|\cdot\log(1/\delta)}\cdot\log|\mathcal{Q}|}{\varepsilon n}\right)^{1/2}$$

Problem: computation time $\mathrm{poly}(n, |\mathcal{X}|, |\mathcal{Q}|)$.

- Exponential in dimensionality of data and query family.
- Inherent in the worst case (cf. “Complexity of DP”).

Private Mult. Weights & Dual Query

[Hardt-Rothblum `10, Gaboardi-Gallego Arias-Hsu-Roth-Wu `14]

(ε, δ)-DP $M\colon \mathcal{X}^n \to \mathcal{X}^m$ such that $\forall q \in \mathcal{Q},\ q\colon \mathcal{X} \to [0,1]$

$$\left|\frac{1}{n}\sum_{i=1}^{n} q(x_i) - \frac{1}{m}\sum_{i=1}^{m} q(M(x)_i)\right| \leq O\left(\frac{\sqrt{\log|\mathcal{X}|\cdot\log(1/\delta)}\cdot\log|\mathcal{Q}|}{\varepsilon n}\right)^{1/2}$$

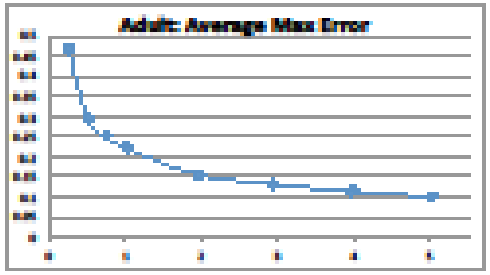
Problem: computation time $\mathrm{poly}(n, |\mathcal{X}|, |\mathcal{Q}|)$.

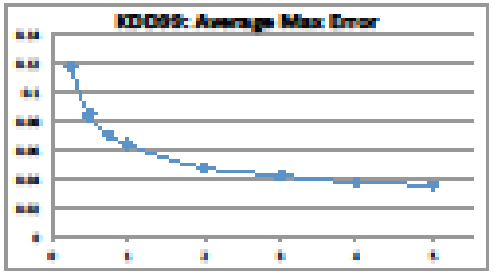
- Exponential in dimensionality of data and query family.
- Inherent in the worst case (cf. "Complexity of DP").

DualQuery:

- Use integer programming get heuristic runtime $\mathrm{poly}(n, \log|\mathcal{X}|, |\mathcal{Q}|)$.
- Privacy doesn't depend on success of heuristic.
- Proven accuracy a bit worse (exponent 1/3 instead of ½).

DualQuery Experiments I





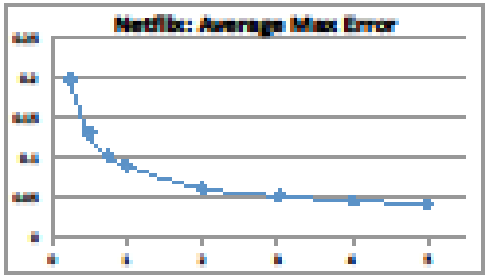


Figure 2: Average max error of $(\varepsilon, 0.001)$-private DualQuery on 500,000 3-way marginals versus ε.

DualQuery Experiments II

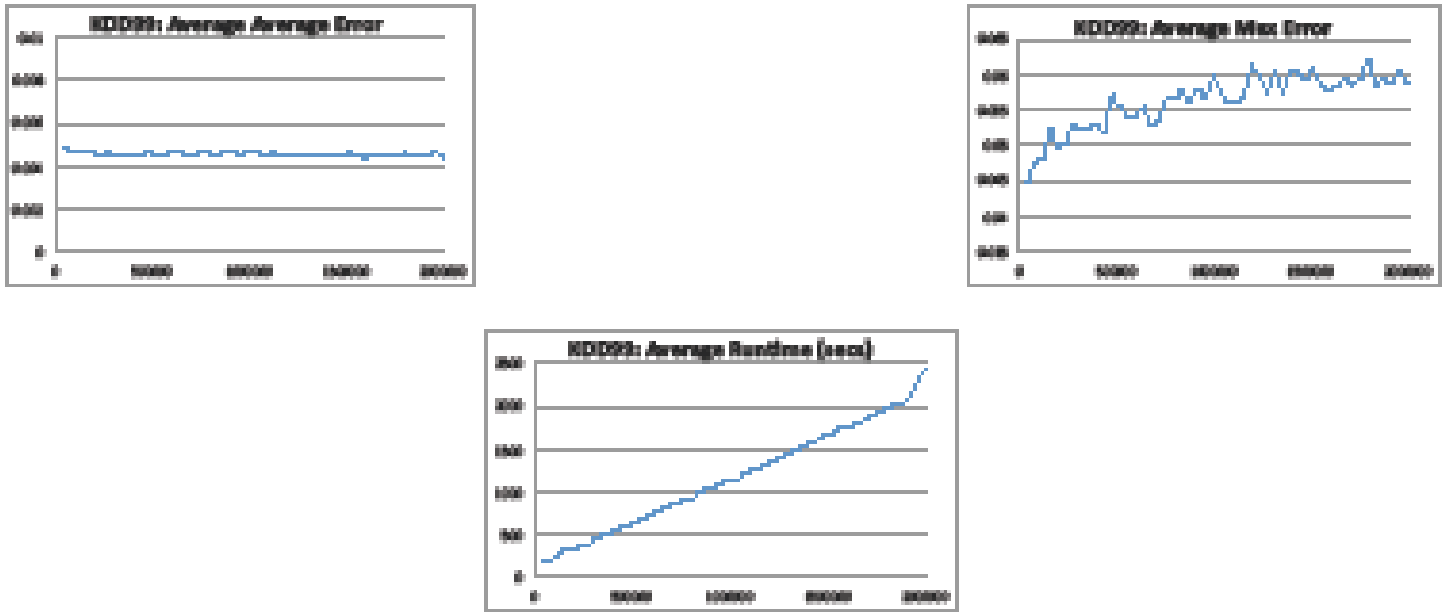
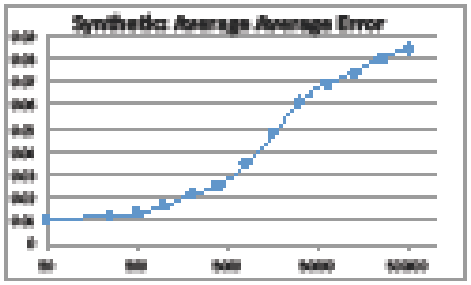
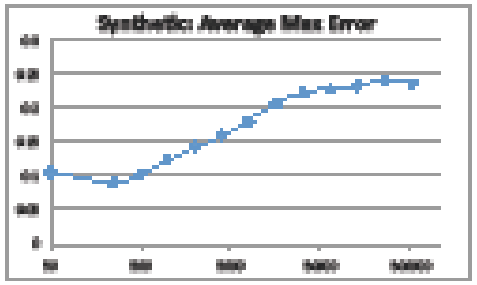


Figure 3: Error and runtime of $(1, 0.001)$-private DualQuery on KDD99 versus number of queries.

DualQuery Experiments III





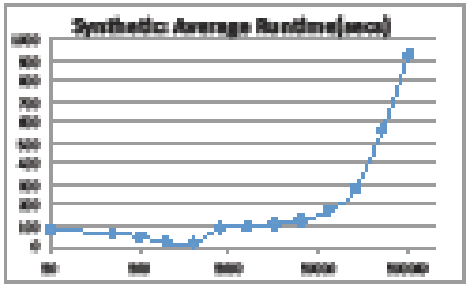


Figure 4: Error and runtime of $(1, 0.001)$-private DualQuery on 100,000 3-way marginal queries versus number of attributes.

See also: Hay et al. “Principled Evaluation of DP Algorithms using DPBench” and https://www.dpcomp.org/