Differential Privacy and Discrepancy on Shortest Paths

Jie Gao

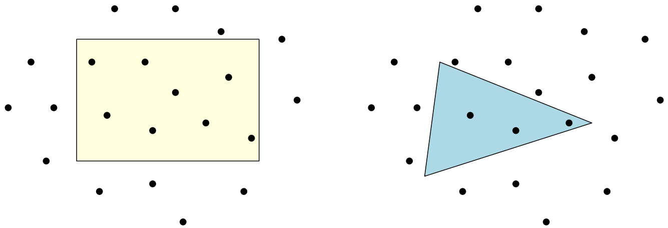
Rutgers University
http://sites.rutgers.edu/jie-gao

September 25th 2024

Classical Range Query Problems

Given points in $\mathbb{R}^d$, report the number of points inside

- Orthogonal ranges: rectilinear boxes in $\mathbb{R}^d$.
- Simplex ranges: d-dimensional simplex (e.g., a triangle in 2D).

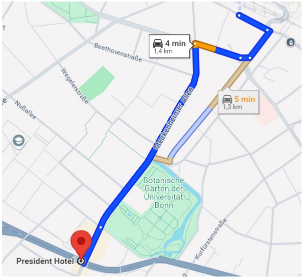


Range Query along Shortest Paths

Given a weighted graph $G = (V, E)$,

- Query ranges = shortest paths $P(s, t)$ on G, $\forall s, t \in V$.
- Edges also carry "sensor readings" that are sensitive and need to be protected with differential privacy guarantee.

Goal: report the sum of sensor readings along a query range $P(s, t)$.



Outline

- **Review of differential privacy**
- 1D range query: Input perturbation vs. output perturbation
- Range query along shortest paths: upper bound
- Lower bound: discrepancy theory
- Open problems

Differential Privacy

[Dwork 06] A randomized range query response mechanism M is ε-differentially private if for any two adjacent datasets D and D' (i.e., differ by ℓ_1 norm of one), for any range $R \in \mathcal{R}$ and any measurable subset $H \in \mathrm{Range}(M)$,

$$\Pr[M_D(R) \in H] \leq e^{\varepsilon} cdot \Pr[M_{D'}(R) \in H].$$

Differential Privacy

[Dwork 06] A randomized range query response mechanism M is ε-differentially private if for any two adjacent datasets D and D' (i.e., differ by ℓ_1 norm of one), for any range $R \in \mathcal{R}$ and any measurable subset $H \in \mathrm{Range}(M)$,

$$\Pr[M_D(R) \in H] \leq e^{\varepsilon} cdot \Pr[M_{D'}(R) \in H].$$

(ε, δ)-differential privacy:

$$\Pr[M_D(R) \in H] \leq e^{\varepsilon} \cdot \Pr[M_{D'}(R) \in H] + \delta.$$

$\delta = 0$: pure-DP; $\delta \neq 0$, approximate-DP.

Why is Differential Privacy a Popular Model?

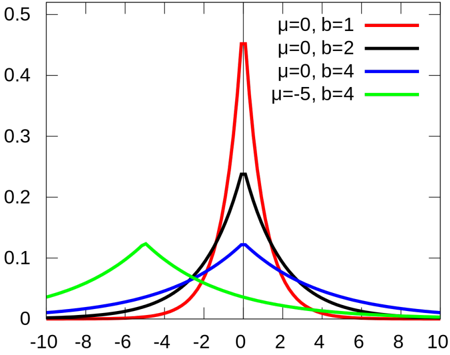
- Post processing of perturbed data does not damage privacy.

Why is Differential Privacy a Popular Model?

- Post processing of perturbed data does not damage privacy.
- Composition (simple): M_1 with ε_1-DP, and M_2 with ε_2-DP, then (M_1, M_2) is $(\varepsilon_1 + \varepsilon_2)$-DP.

Laplace Mechanism

Laplace mechanism: add noise with distribution Lap(b), and its probability density is given as: $\text{Lap}[x|b] = \frac{1}{2b}\exp(-\frac{|x|}{b})$.



Laplace Mechanism

The level of noise is usually determined by sensitivity.

Laplace Mechanism

The level of noise is usually determined by sensitivity.

The sensitivity of a function f, written as Δf, is the largest possible difference in the output of f between any pair of adjacent databases:

$$\max_{(D,D')} |f(D) - f(D')|.$$

Laplace Mechanism

The level of noise is usually determined by sensitivity.

The sensitivity of a function f, written as Δf, is the largest possible difference in the output of f between any pair of adjacent databases:

$$\max_{(D,D')} |f(D) - f(D')|.$$

Example: f as the average employee salary.

Laplace Mechanism satisfies DP

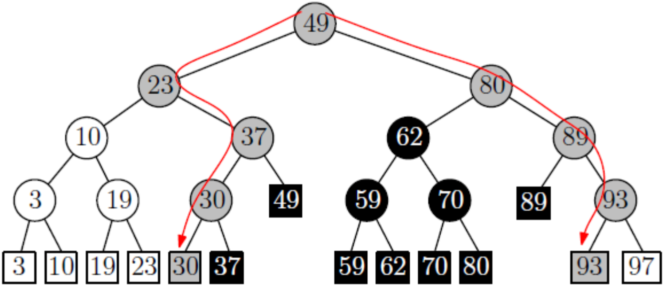
To achieve ε-differential privacy, adding noise $z \sim \text{Lap}(\Delta f / \varepsilon)$ suffices.

Laplace Mechanism satisfies DP

To achieve ε-differential privacy, adding noise $z \sim \text{Lap}(\Delta f/\varepsilon)$ suffices.

$$
\begin{aligned}
&\frac{\Pr[z + f(D) = x]}{\Pr[z' + f(D') = x]} \\
=&\frac{\exp(-|x - f(D)|/b)}{\exp(-|x - f(D')|/b)}, b = \frac{\Delta f}{\varepsilon} \\
\leq& \exp(\varepsilon \cdot \frac{|x - f(D')| - |x - f(D)|}{\Delta f}) \\
\leq& e^{\varepsilon}
\end{aligned}
$$

1D Range Tree



Two types of DP mechanisms:

- Input perturbation: add noise to each input element.
- Output perturbation: add noise to the query results.

Input Perturbation

Publish data with iid noise $\sim \mathrm{Lap}(1/\varepsilon)$ on each element.

Input Perturbation

Publish data with iid noise $\sim \text{Lap}(1/\varepsilon)$ on each element.

- ε-DP.
- Answer queries on perturbed data in the normal way. $\rightarrow$ Post-processing.

Input Perturbation

Publish data with iid noise $\sim \text{Lap}(1/\varepsilon)$ on each element.

- ε-DP.
- Answer queries on perturbed data in the normal way. $\rightarrow$ Post-processing.

What is the error magnitude of a query on *n* elements?

Sum of Independent Laplace Variables

[CCS'11] Suppose $\gamma_i \sim \mathsf{Lap}(b_i)$ and $Y = \sum_i \gamma_i$. Then, with $0 < \delta < 1$, $\Pr[|Y| = O(\sqrt{\sum_i b_i^2} \log(1/\delta))] \geq 1 - \delta$.

Sum of Independent Laplace Variables

[CCS'11] Suppose $\gamma_i \sim \text{Lap}(b_i)$ and $Y = \sum_i \gamma_i$. Then, with $0 < \delta < 1$, $\Pr[|Y| = O(\sqrt{\sum_i b_i^2} \log(1/\delta))] \geq 1 - \delta$.

If a query range consists of n elements, where each is added an independent noise from $\text{Lap}(1/\varepsilon)$, then the total error $\sim O(\frac{1}{\varepsilon}\sqrt{n} \log \frac{1}{\delta})$ with probability $1 - \delta$.

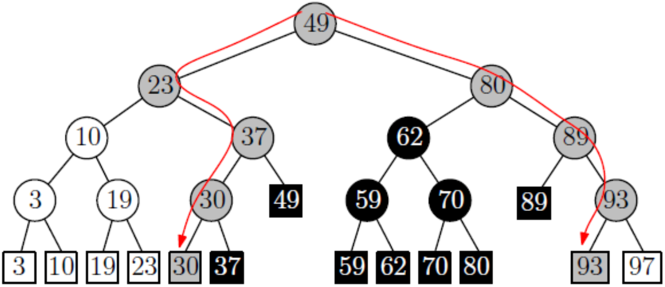
Output Perturbation

Answer a query with a fresh noise $\sim \mathrm{Lap}(1/\varepsilon)$.

- If an element is involved in m queries, then we have $(m\varepsilon)$-DP.
- Or we enforce ε-DP, query error $\sim O(m/\varepsilon)$.
- m could be $\sim n^2$.

Combining Input and Output Perturbation

[CCS'11] Add iid noise $\sim \text{Lap}(\log n/\varepsilon)$ on each node of the range tree.



Error: sum up $O(\log n)$ iid noise, each $\sim \text{Lap}(\log n/\varepsilon) \Rightarrow O(\log^{1.5} n/\varepsilon)$.

Outline

- Review of differential privacy
- 1D range query: Input perturbation vs. output perturbation
- **Range query along shortest paths: upper bound**
- Lower bound: discrepancy theory
- Open problems

Range Query along Shortest Paths

Input perturbation:

- Add iid noise $\sim \text{Lap}(1/\varepsilon)$ to each edge value.
- Query error?

Range Query along Shortest Paths

Input perturbation:

- Add iid noise $\sim \text{Lap}(1/\varepsilon)$ to each edge value.
- Query error? $O(n/\varepsilon)$.

Range Query along Shortest Paths

Input perturbation:

- Add iid noise $\sim \text{Lap}(1/\varepsilon)$ to each edge value.
- Query error? $O(n/\varepsilon)$.

Output perturbation:

- Add iid $\sim \text{Lap}(Y/\varepsilon)$ to each query output.

Range Query along Shortest Paths

Input perturbation:

- Add iid noise $\sim \text{Lap}(1/\varepsilon)$ to each edge value.
- Query error? $O(n/\varepsilon)$.

Output perturbation:

- Add iid $\sim \text{Lap}(Y/\varepsilon)$ to each query output.
- What is Y? – the number of queries that may contain one vertex, $Y = \Theta(n^2)$.
- Query error $O(n^2/\varepsilon)$.

Use Canonical Paths [Deng, G, Upadhyay, Wang'23]

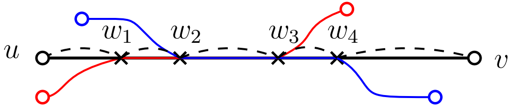
- Randomly sample s vertices S, build shortest paths between all pairs in S.

Use Canonical Paths [Deng, G, Upadhyay, Wang'23]

- Randomly sample s vertices S, build shortest paths between all pairs in S.
- Take intersection of all $O(s^2)$ paths. wlog assume these paths are unique shortest paths.

Use Canonical Paths [Deng, G, Upadhyay, Wang'23]

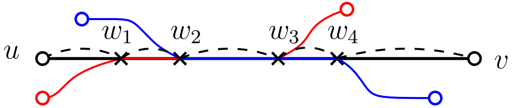
- Randomly sample s vertices S, build shortest paths between all pairs in S.
- Take intersection of all $O(s^2)$ paths. wlog assume these paths are unique shortest paths.



- Such intersections partition a single path into $O(s^2)$ canonical segments.

Use Canonical Paths [Deng, G, Upadhyay, Wang'23]

- Randomly sample s vertices S, build shortest paths between all pairs in S.
- Take intersection of all $O(s^2)$ paths. wlog assume these paths are unique shortest paths.



- Such intersections partition a single path into $O(s^2)$ canonical segments.

Claim: any two canonical segments are edge disjoint.

Use Canonical Paths

DP mechanism:

- Input perturbation on each edge: Lap$(2/\varepsilon)$

Use Canonical Paths

DP mechanism:

- Input perturbation on each edge: $\text{Lap}(2/\varepsilon)$
- Output perturbation on each canonical segment: $\text{Lap}(2/\varepsilon) \Rightarrow$ each edge may appear in at most one canonical segment.

Use Canonical Paths

DP mechanism:

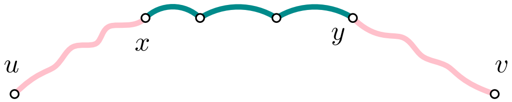
- Input perturbation on each edge: $\mathrm{Lap}(2/\varepsilon)$
- Output perturbation on each canonical segment: $\mathrm{Lap}(2/\varepsilon) \Rightarrow$ each edge may appear in at most one canonical segment.

Adding up, we have $\varepsilon/2 + \varepsilon/2 = \varepsilon$-DP.

Use Canonical Paths

Error analysis: fix a shortest path $P(u, v)$. Along $P(u, v)$

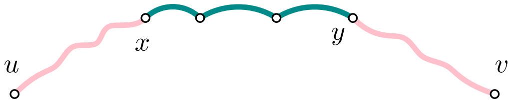
- # vertices/edges before reaching the first vertex x in S: $\tilde{O}(n/s)$.
- Take perturbed values from $O(s^2)$ canonical segments until the last vertex y on P.
- From u to x and from y to v use input perturbation.



Use Canonical Paths

Error analysis: fix a shortest path $P(u, v)$. Along $P(u, v)$

- $\#$ vertices/edges before reaching the first vertex x in S: $\tilde{O}(n/s)$.
- Take perturbed values from $O(s^2)$ canonical segments until the last vertex y on P.
- From u to x and from y to v use input perturbation.



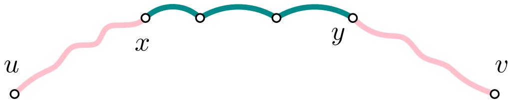
Total error:

$$\tilde{O}\left(\frac{1}{\varepsilon} \cdot \sqrt{\frac{n}{s} + s^2}\right)$$

Use Canonical Paths

Error analysis: fix a shortest path $P(u, v)$. Along $P(u, v)$

- # vertices/edges before reaching the first vertex x in S: $\tilde{O}(n/s)$.
- Take perturbed values from $O(s^2)$ canonical segments until the last vertex y on P.
- From u to x and from y to v use input perturbation.



Total error:

$$\tilde{O}(\frac{1}{\varepsilon} \cdot \sqrt{\frac{n}{s} + s^2})$$

Take $s = n^{1/3}$ we get error of $\tilde{O}(n^{1/3}/\varepsilon)$.

Improve the upper bound to $\tilde{O}(n^{1/4})$

[Ashvinkumar, Bernstein, Deng, G, Wein'24] Process shortest paths in an order. Vertices on processed paths are 'frozen'.

- Take P with max # unfrozen vertices.
- Apply DP as in the 1D range query along P, only on new vertices, with $\text{Lap}(2 \log n/\varepsilon)$.
- All remaining edges w/ input perturbation $\text{Lap}(2/\varepsilon)$

Improve the upper bound to $\tilde{O}(n^{1/4})$

[Ashvinkumar, Bernstein, Deng, G, Wein'24] Process shortest paths in an order. Vertices on processed paths are 'frozen'.

- Take P with max # unfrozen vertices.
- Apply DP as in the 1D range query along P, only on new vertices, with $\mathrm{Lap}(2\log n/\varepsilon)$.
- All remaining edges w/ input perturbation $\mathrm{Lap}(2/\varepsilon)$

Claim: along any shortest path, we have

- at most $\sqrt{n}$ 'frozen' segments.
- at most $\sqrt{n}$ edges between frozen segments.

Improve the upper bound to $\tilde{O}(n^{1/4})$

[Ashvinkumar, Bernstein, Deng, G, Wein'24] Process shortest paths in an order. Vertices on processed paths are 'frozen'.

- Take P with max # unfrozen vertices.
- Apply DP as in the 1D range query along P, only on new vertices, with $\mathrm{Lap}(2\log n/\varepsilon)$.
- All remaining edges w/ input perturbation $\mathrm{Lap}(2/\varepsilon)$

Claim: along any shortest path, we have

- at most $\sqrt{n}$ 'frozen' segments.
- at most $\sqrt{n}$ edges between frozen segments.

Total error:

$$\tilde{O}(\frac{1}{\varepsilon} \cdot n^{1/4})$$

Outline

- Review of differential privacy
- 1D range query: Input perturbation vs. output perturbation
- Range query along shortest paths: upper bound
- **Lower bound: discrepancy theory**
- Open problems

Lower Bound by Discrepancy Theory

Incidence matrix M with $\binom{n}{2}$ rows (paths) and n columns (vertices). Multiply M with a vector x of $\{+1,-1\}^n$.

$$\begin{pmatrix} 1 & 0 & \cdots & \\ \cdots & & & \\ \cdots & & & \\ \cdots & & & \end{pmatrix} \cdot \begin{bmatrix} +1 \\ -1 \\ \vdots \\ +1 \end{bmatrix}$$

The minimum L_∞ norm over all vector x, vertex discrepancy, is a lower bound on DP-error. [Muthukrishnan, Nikolov'12]

If edges carry sensitive values: take m edges as columns – edge discrepancy. We consider vertex discrepancy first.

Primal Shatter Function of Shortest Paths

Primal shatter function $\pi_{\mathcal{R}}(s)$: maximum number of distinct sets in $\{A \cap S \mid S \in \mathcal{S}\}$ for some $A \subseteq X$ such that $|A| = s$.

Primal Shatter Function of Shortest Paths

Primal shatter function $\pi_{\mathcal{R}}(s)$: maximum number of distinct sets in $\{A \cap S \mid S \in \mathcal{S}\}$ for some $A \subseteq X$ such that $|A| = s$.

- For both undirected graphs and directed graphs: $\pi_{\mathcal{R}}(s) = O(s^2)$

Primal Shatter Function of Shortest Paths

Primal shatter function $\pi_{\mathcal{R}}(s)$: maximum number of distinct sets in $\{A \cap S \mid S \in \mathcal{S}\}$ for some $A \subseteq X$ such that $|A| = s$.

- For both undirected graphs and directed graphs: $\pi_{\mathcal{R}}(s) = O(s^2)$

[Matousek'95] vertex discrepancy is $O(n^{1/2-1/(2d)}) = O(n^{1/4})$.

Discrepancy of Path Systems

[Bodwin, Deng, G, Hoppenworth, Upadhyay, Wang'24]
For a general set of $O(n)$ paths, the discrepancy can be $\Omega(\sqrt{n})$.

Discrepancy of Path Systems

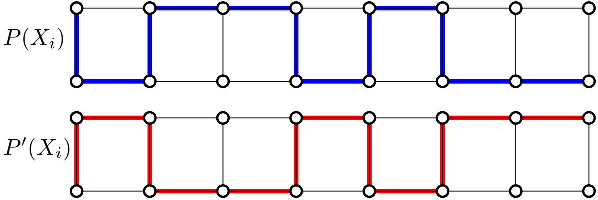
[Bodwin, Deng, G, Hoppenworth, Upadhyay, Wang'24]
For a general set of $O(n)$ paths, the discrepancy can be $\Omega(\sqrt{n})$.
The Hadamard matrix H_n.

$$H_8 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

The discrepancy of $\frac{1}{2}(H+J)$ is $\Omega(\sqrt{n})$, J is a all-1 matrix.

Discrepancy of Path Systems

'Embed' the Hadamard matrix on a $2 \times n$ grid: for a row:
$X_i = (1, 1, 0, 1, 1, 1, 0, 0)$



In addition, add P and P' to be the top/bottom path.

Discrepancy of Point-Line System: Lower Bound $\Omega(n^{1/6})$

Erdös point-line system: n points, n lines with

- each point staying on $\Theta(n^{1/3})$ lines;
- each line through $\Theta(n^{1/3})$ points.

Hereditary discrepancy of the point-line incidence matrix is $\Omega(n^{1/6})$ (Apply the trace bouund).

Discrepancy of Point-Line System: Lower Bound $\Omega(n^{1/6})$

Erdös point-line system: n points, n lines with

- each point staying on $\Theta(n^{1/3})$ lines;
- each line through $\Theta(n^{1/3})$ points.

Hereditary discrepancy of the point-line incidence matrix is $\Omega(n^{1/6})$ (Apply the trace bouund).

Take edge weights as L_2 distances $\Rightarrow$ Every line is a shortest path.

Discrepancy of Shortest Paths: Lower Bound $\Omega(n^{1/4})$

[Bodwin, Deng, G, Hoppenworth, Upadhyay, Wang'24] Adapt the construction [Bodwin, Hoppenworth'23] for hopset lower bound here:

- Start from point-line incidence system.
- Shift the points to allow shortest paths to have longer 'overlap'.
- Planarize it by adding crossing vertices & adjusting edge weights.
- Sparse graph: $O(n \log^6 n)$ nodes

Discrepancy of Shortest Paths: Lower Bound $\Omega(n^{1/4})$

[Bodwin, Deng, G, Hoppenworth, Upadhyay, Wang'24] Adapt the construction [Bodwin, Hoppenworth'23] for hopset lower bound here:

- Start from point-line incidence system.
- Shift the points to allow shortest paths to have longer 'overlap'.
- Planarize it by adding crossing vertices & adjusting edge weights.
- Sparse graph: $O(n \log^6 n)$ nodes

Apply trace bound to get $\Omega(n^{1/4})$.

Open Problem #1

What is the edge discrepancy for shortest paths in a directed graph?

- $O(m^{1/4})$: primal shatter function $O(s^2)$.
- $O(D^{1/2}) = O(n^{1/2})$ with diameter D: by random coloring.
- DAG: $O(n^{1/4})$ – shortest paths are consistent & constructive upper bound.
- Lower bound $\Omega(n^{1/4})$.

Open Problem #1

What is the edge discrepancy for shortest paths in a directed graph?

- $O(m^{1/4})$: primal shatter function $O(s^2)$.
- $O(D^{1/2}) = O(n^{1/2})$ with diameter D: by random coloring.
- DAG: $O(n^{1/4})$ – shortest paths are consistent & constructive upper bound.
- Lower bound $\Omega(n^{1/4})$.

A stronger lower bound shall be non-DAG, non-sparse with large diameter.

Open Problem #2

Publish differentially private all-pairs shortest distances: graph topology is public, edge weight is sensitive.

- Upper bound on error $O(\sqrt{n})$. [Chen, Ghazi, Kumar, Manurangsi, Narayanan, Nelson, Xu'23, Fan, Li, Li'23]
- Our discrepancy lower bound $\Omega(n^{1/4})$ applies.

One cannot use the shortest paths to design the DP mechanism.