Differentially Private Range Query on Shortest Paths

Jie Gao

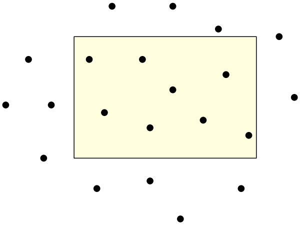
Rutgers University
http://sites.rutgers.edu/jie-gao

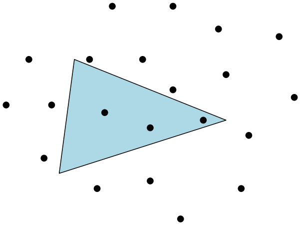
December 12 2023

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 on 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".

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

Range Query on Shortest Paths

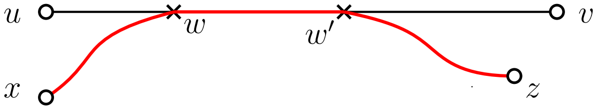
Given a weighted graph $G = (V, E)$,

- Query ranges = shortest paths $P(s, t)$ on G, $\forall s, t \in V$.
- Edges also carry "sensor readings".

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

Assumptions

1. The shortest paths are 'consistent' – any two shortest paths intersect at a contiguous subpath.



Range Query on Shortest Paths

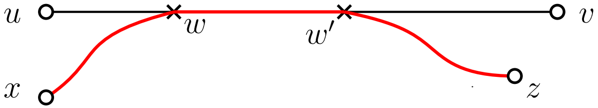
Given a weighted graph $G = (V, E)$,

- Query ranges = shortest paths $P(s, t)$ on G, $\forall s, t \in V$.
- Edges also carry "sensor readings".

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

Assumptions

1. The shortest paths are 'consistent' – any two shortest paths intersect at a contiguous subpath.



2. The sensor readings are sensitive and need to be protected with privacy guarantee.

Plan

- **Review of differential privacy**
- 1D range query: Input perturbation vs. output perturbation
- Combining input perturbation vs. output perturbation
- Range query on shortest paths
- Connection to VC-dimension and discrepancy theory

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 $H \in \mathrm{Range}(M)$,

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

$(\varepsilon,\ \delta)$-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?

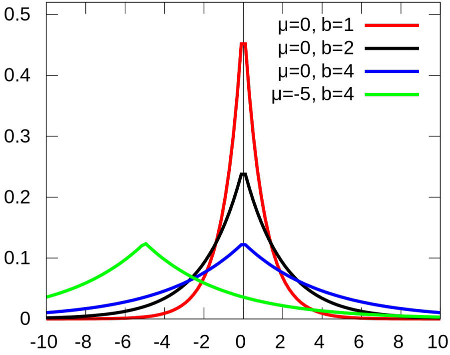
- Post processing of perturbed data does not damage privacy.

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 in terms of *sensitivity*.

Laplace Mechanism

The level of noise is usually determined in terms of *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 in terms of *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 $z \sim \mathrm{Lap}(\Delta f/\varepsilon)$ suffices.

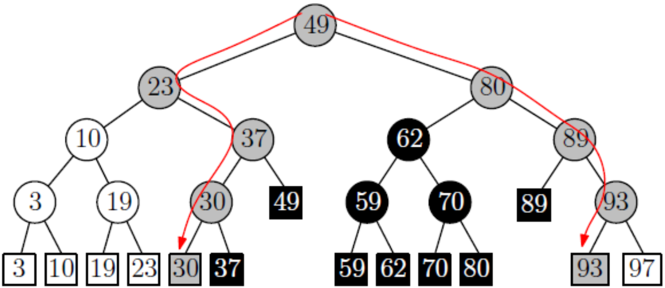
Laplace Mechanism Satisfies DP

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

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

1D Range Tree

Build a binary search tree. Run two queries for the boundary of $[25, 90]$. Take points in between.



Input Perturbation

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

Input Perturbation

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

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

Input Perturbation

Publish data with iid noise $\sim \mathrm{Lap}(1/\varepsilon)$ 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 $\gamma_i \sim \mathsf{Lap}(b_i)$ and $Y = \sum_i \gamma_i$. Then, with $0 < \delta < 1$, $\Pr[|Y| = O(\sqrt{\sum_i b_i^2} \log(1/\delta))] \geq 1 - \delta$.

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

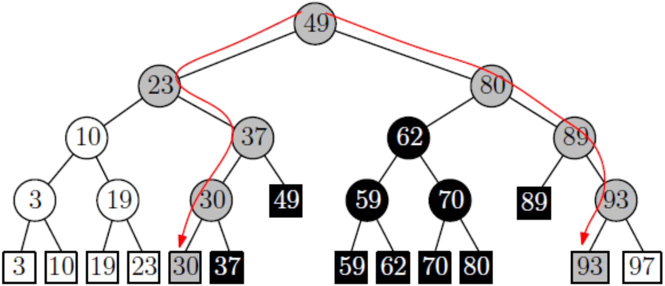
Output Perturbation

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

- If an element is involved in m queries, then we have $(m\varepsilon)$-DP.
- Or we take $\varepsilon' = m\varepsilon$, query error $\sim O(m/\varepsilon)$.

Combining Input and Output Perturbation

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



Query error? Sum up $O(\log n)$ independent noise, each $\sim \text{Lap}(\log n/\varepsilon) \Rightarrow O(\log^{1.5} n/\varepsilon)$.

Plan

- Review of differential privacy
- 1D range query: Input perturbation vs. output perturbation
- Combining input and output perturbation
- **Range query on shortest paths**
- Connection to VC-dimension and discrepancy theory

Range Query on Shortest Paths

Input perturbation:

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

Range Query on Shortest Paths

Input perturbation:

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

Range Query on Shortest Paths

Input perturbation:

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

Output perturbation:

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

Range Query on 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

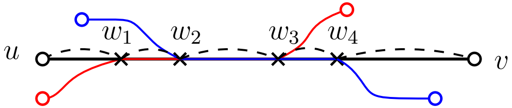
- Randomly sample s vertices S, build shortest paths between all pairs in S.

Use Canonical Paths

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

Use Canonical Paths

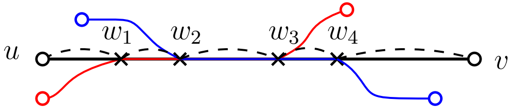
- Randomly sample s vertices S, build shortest paths between all pairs in S.
- Take intersection of all $O(s^2)$ paths.



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

Use Canonical Paths

- Randomly sample s vertices S, build shortest paths between all pairs in S.
- Take intersection of all $O(s^2)$ 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 pertuabtion on each edge value: $\mathrm{Lap}(2/\varepsilon)$

Use Canonical Paths

DP mechanism:

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

Use Canonical Paths

DP mechanism:

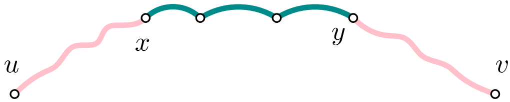
- Input pertuabtion on each edge value: Lap($2/\varepsilon$)
- Output perturbation on canonical segments: Lap($2/\varepsilon$) $\Rightarrow$ each edge may appear in at most one canonical segment.

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

Use Canonical Paths

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

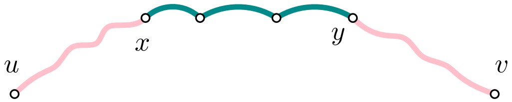
- $\#$ vertices before reaching the first vertex x in S: $\tilde{O}(n/s)$.
- Take perturbed values from $O(s^2)$ canonical segments until the last vertex y on P.
- From u to x and from y to v use input perturbation.



Use Canonical Paths

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

- $\#$ vertices 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.



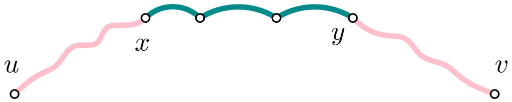
Total error:

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

Use Canonical Paths

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

- $\#$ vertices 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 $\tilde{O}(n^{1/3}/\varepsilon)$.

Approximate-DP; Better Error Bounds

Gaussian noise $N(0, \sigma^2)$ with $\sigma^2 \approx \frac{(\Delta f)^2}{\varepsilon^2} \log \frac{1}{\delta}$ gives (ε, δ)-DP.

Approximate-DP; Better Error Bounds

Gaussian noise $N(0, \sigma^2)$ with $\sigma^2 \approx \frac{(\Delta f)^2}{\varepsilon^2} \log \frac{1}{\delta}$ gives (ε, δ)-DP.

Strong composition: With k (ε, δ)-DP mechanisms $\Rightarrow$ (ε', δ')-DP with $\varepsilon' \approx \varepsilon\sqrt{k}$ and $\delta' \approx k\delta$.

Approximate-DP; Better Error Bounds

Gaussian noise $N(0, \sigma^2)$ with $\sigma^2 \approx \frac{(\Delta f)^2}{\varepsilon^2} \log \frac{1}{\delta}$ gives (ε, δ)-DP.

Strong composition: With k (ε, δ)-DP mechanisms $\Rightarrow$ (ε', δ')-DP with $\varepsilon' \approx \varepsilon\sqrt{k}$ and $\delta' \approx k\delta$.

Algorithm: 1. Add Gaussian noise for each edge with $(\varepsilon/2, \delta/2)$-DP.

Approximate-DP; Better Error Bounds

Gaussian noise $N(0, \sigma^2)$ with $\sigma^2 \approx \frac{(\Delta f)^2}{\varepsilon^2} \log \frac{1}{\delta}$ gives (ε, δ)-DP.

Strong composition: With k (ε, δ)-DP mechanisms $\Rightarrow$ (ε', δ')-DP with $\varepsilon' \approx \varepsilon\sqrt{k}$ and $\delta' \approx k\delta$.

Algorithm: 1. Add Gaussian noise for each edge with $(\varepsilon/2, \delta/2)$-DP.

2. Randomly sample s vertices S, $\forall s \in S$

- Build shortest path tree.
- On each tree, run heavy-light decomposition.
- Add Gaussian noise for each heavy path. $\Rightarrow$ each tree gives $(\varepsilon/\sqrt{s}, \delta/s)$-DP.

Approximate-DP; Better Error Bounds

Error analysis: summation of

- $O(n/s)$ Gaussian noises $\approx O_{\varepsilon,\delta}(1)$, and
- $O(\log n)$ heavy paths each of noise $\approx O_{\varepsilon,\delta}(\sqrt{s})$, and
- $O(\log n)$ light edges of noises $\approx O_{\varepsilon,\delta}(1)$

Total error:

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

Take $s = n^{1/2}$ we get error of $\tilde{O}_{\varepsilon,\delta}(n^{1/4})$.

Approximate-DP; Better Error Bounds

Error analysis: summation of

- $O(n/s)$ Gaussian noises $\approx O_{\varepsilon,\delta}(1)$, and
- $O(\log n)$ heavy paths each of noise $\approx O_{\varepsilon,\delta}(\sqrt{s})$, and
- $O(\log n)$ light edges of noises $\approx O_{\varepsilon,\delta}(1)$

Total error:

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

Take $s = n^{1/2}$ we get error of $\tilde{O}_{\varepsilon,\delta}(n^{1/4})$.

Q: Improve further? Lower bound?

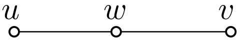
Plan

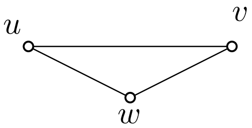
- Review of differential privacy
- 1D range query: Input perturbation vs. output perturbation
- Combining input perturbation vs. output perturbation
- Range query on shortest paths
- **Connection to VC-dimension and discrepancy theory**

VC-dimension of Shortest Paths

Consistent shortest paths have low VC-dimension:

- Undirected graph: VC-dimension $=2$ [ADFGW 11][TSP 11]

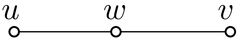


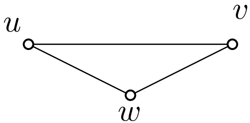


VC-dimension of Shortest Paths

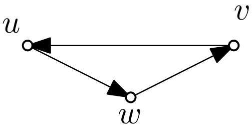
Consistent shortest paths have low VC-dimension:

- Undirected graph: VC-dimension $=2$ [ADFGW 11][TSP 11]





- Directed graph: VC-dimension $= 3$ [FNS 14]



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: $\pi_{\mathcal{R}}(s) = O(s^2)$

[Muthukrishnan and Nikolov 12]: Range query with primal function $O(s^d)$ admits (ε, δ)-DP algorithm with error $O_{\varepsilon,\delta}(m^{1/2-1/(2d)})$, where m is the size of the ground set.

Discrepancy of Shortest Paths

If we assign colors $\{+1, -1\}$ to vertices (or edges) of a graph, what is the discrepancy of consistent shortest paths in a graph?

[Chen et al 23]: (hereditary) discrepancy is a lower bound of the approx-DP error.

Discrepancy of Shortest Paths

If we assign colors $\{+1, -1\}$ to vertices (or edges) of a graph, what is the discrepancy of consistent shortest paths in a graph?

[Chen et al 23]: (hereditary) discrepancy is a lower bound of the approx-DP error.

Erdös point-line system: n points on 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 $\Omega(n^{1/6})$. – Edge weights as L_2 distances $\Rightarrow$ shortest paths discrepancy.

Discrepancy of Shortest Paths

If we assign colors $\{+1, -1\}$ to vertices (or edges) of a graph, what is the discrepancy of consistent shortest paths in a graph?

[Chen et al 23]: (hereditary) discrepancy is a lower bound of the approx-DP error.

Erdös point-line system: n points on 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 $\Omega(n^{1/6})$. – Edge weights as L_2 distances $\Rightarrow$ shortest paths discrepancy.

New results: discrepancy lower bound $\Omega(n^{1/4})$ on shortest paths discrepancy.

Acknowledgement

- Rutgers: Chengyuan Deng, Jalaj Upadhyay, Chen Wang
- Michigan: Greg Bodwin, Gary Hoppenworth

Questions and Comments?