On Differentially Private U statistics and Application to Random Geometric Networks

Purnamrita Sarkar
(With Kamalika Chaudhuri, Po-Ling Loh, Shourya Pandey)

IMSI 2026

Jan 15, 2026

Roadmap

- Differential Privacy
- U statistics
- Random graphs

Differential Privacy

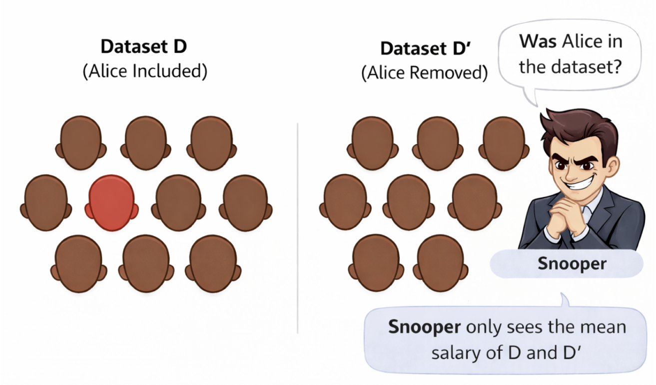


Figure: A simple picture - thanks ChatGPT!

Differential Privacy

Consider a dataset $X_1, X_2, \ldots, X_n$ and another dataset $Y_1, Y_2, \ldots, Y_n$ such that $X_i = Y_i$ except possibly at one index. Then, differential privacy means

$$\tilde{h}(X_1, X_2, \ldots, X_n) \stackrel{d}{\approx} \tilde{h}(Y_1, Y_2, \ldots, Y_n).$$

The output distribution is not affected much by changing any single data point.

Differential Privacy

(ϵ, δ)-differential privacy (Dwork and Roth, 2014)

An algorithm $\tilde{h} : \mathcal{X}^n \to \mathbb{R}$ is (ϵ, δ)-differentially private if for any index $i \in [n]$ and adjacent datasets D and D',

$$\mathbb{P}(\tilde{h}(D) \in S) \leq e^{\epsilon} \cdot \mathbb{P}(\tilde{h}(D') \in S) + \delta.$$

Laplace mechanism

More generally, we have a way to privatize algorithms that change in value by at most Δ between adjacent datasets.

Laplace Mechanism

Let $f : \mathcal{X}^n \to \mathbb{R}$ be a deterministic estimator such that for any adjacent datasets D and D', $|f(D) - f(D')| \leq \Delta$. Then, the estimator

$$\tilde{f}(D) = f(D) + \frac{\Delta}{\epsilon} \cdot Z,$$

where $Z \sim \mathrm{Lap}\,(1)$, is $(\epsilon, 0)$-differentially private.

We call Δ the global sensitivity of the estimator f.

Global sensitivity is worst case

- Global sensitivity can be viewed as a worst case bound
- Does not always capture the true behavior of the function
- Wouldn't it be nice if we can use something more **local**?

Local sensitivity

How sensitive is this statistic near my actual dataset, not in some pathological worst case?

Global sensitivity is worst case

- Define local sensitivity $LS(D) := \max_{D'} |f(D) - f(D')|$
- But this leaks information!

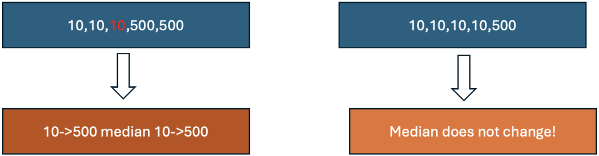


Figure: Why local sensitivity itself leaks privacy

Global sensitivity is worst case

Not global, not local

How about we find an upper bound on the local sensitivity, that does not change too much between neighboring datasets?

Smoothed Sensitivity!

Smoothed sensitivity

Smoothed sensitivity (Nissim et al., 2007)

Given a function $f : \mathcal{X}^n \to \mathbb{R}$, we call $SS : \mathcal{X}^n \to \mathbb{R}$ a β-smooth upper bound on the local sensitivity of f if

1. SS is an upper bound on the local sensitivity of f.

$$\max_{D' \sim D} |f(D) - f(D')| \leq SS(D) \;\; \forall D \in \mathcal{X}^n.$$

2. SS is β-smooth. For any adjacent datasets $D, D' \in \mathcal{X}^n$,

$$SS(D) \leq e^{\beta} \cdot SS(D').$$

Smoothed sensitivity

Smoothed Sensitivity Mechanism

Let $\epsilon, \delta \in (0, 1)$. Let $f : \mathcal{X}^n \to \mathbb{R}$ be a deterministic estimator with a β-smooth upper bound SS on its local sensitivity, where $\beta \leq \frac{\epsilon}{2\ln(2/\delta)}$. Then, the estimator

$$\tilde{f}(D) = f(D) + \frac{2 \cdot SS(D)}{\epsilon} \cdot Z,$$

where $Z \sim \mathrm{Lap}(1)$, is (ϵ, δ)-differentially private.

Roadmap

- Differential Privacy
- U statistics
- Random graphs

Differentially Private Parameter Estimation

Differentially Private Estimation

Given n independent and identically distributed (IID) samples $X_1, X_2, \ldots, X_n$ from a distribution $\mathcal{D}$ supported on $\mathcal{X}$, and a real-valued function $h : \mathcal{X}^k \to \mathbb{R}$, devise an estimator $\tilde{h}$ that has small MSE

$$\mathbb{E}[(\tilde{h}(X_1, X_2, \ldots, X_n) - \theta)^2],$$

where $\theta := \mathbb{E}_{Y_1,\ldots,Y_k \sim \mathcal{D}}[h(Y_1, Y_2, \ldots, Y_k)]$.

The estimator $\tilde{h}$ should be (ϵ, δ)-differentially private.

U-statistics

U-statistic

Let $h : \mathcal{X}^k \to \mathbb{R}$ be a symmetric function (called the kernel) and $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{D}$. Then,

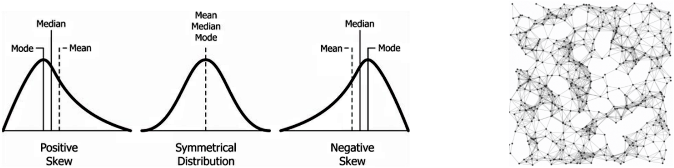
$$U_n := \binom{n}{k}^{-1} \sum_{S \subseteq \binom{[n]}{k}} h(X_S)$$

is the associated U-statistic of order k.

U_n is an unbiased estimator of $\theta = \mathbb{E}_{Y_1,\ldots,Y_k \sim \mathcal{D}}[h(Y_1, \ldots, Y_k)]$.

Examples

- Mean: $\theta = \mathbb{E}[X_1]$, $k = 1$
- Variance $\theta = \mathbb{E}[(X_1 - X_2)^2/2]$
- Symmetry testing:
  $h(X_1, X_2, X_3) = \text{median}(X_1, X_2, X_3) - \text{mean}(X_1, X_2, X_3)$
- Subgraph counts in geometric random graphs



(A) (B)

Figure: (A) Testing for symmetry, (B) Geometric graph

Variance of a U-statistic

Since U_n is unbiased, the error $\mathbb{E}[(U_n - \theta)^2]$ of U_n is equal to

$$\mathrm{Var}(U_n) = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c} \binom{n-c}{k-c} \zeta_c,$$

where ζ_c is the conditional variance

$$\mathrm{Var}(\mathbb{E}[Y_1, Y_2, \ldots, Y_k | Y_1, Y_2, \ldots, Y_c]) = \mathrm{Cov}_{|S \cap S'| = c}(h(X_S), h(X_{S'})).$$

U-statistics: The best among the fairest

In many applications, like hypothesis testing, under the null, the variance is $O(1/n^2)$ and not $O(1/n)$.

How about applying existing methods

- Assume $h(X_1, \dots, X_k) \sim \text{subgaussian}(\tau)$

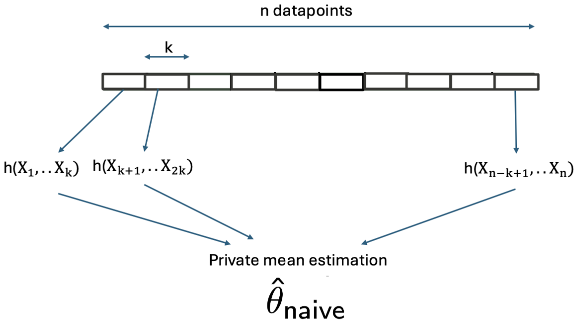


Figure: Applying private mean estimation naively

- Informally, with high probability,

$$|\hat{\theta}_{naive} - \theta| \leq \underbrace{O\left(\sqrt{k/n}\right)}_{\text{Non-private error}} + \underbrace{\tilde{O}\left(k/n\epsilon\right)}_{\text{Error incurred for privacy guarantee}},$$

Is this satisfactory?

- With high probability,

$$|\hat{\theta}_{naive} - \theta| \leq \underbrace{O\left(\sqrt{\frac{k}{n}}\right)}_{\text{Non-private error}} + \underbrace{\tilde{O}\left(\frac{k}{n\epsilon}\right)}_{\text{Error incurred for privacy guarantee}},$$

- Turns out, in many hypothesis tests, under the null, or in the neighborhood of the null, var(U_n) can be in fact be $O(1/n^2)$ (Lee, 2019).
  - The naive estimator's nonprivate error can be an order off in these situations.
- Examples include uniformity testing, subgraph counts in random geometric graphs and many more!

Differentially Private Parameter Estimation

Private U-statistic?

How can we privatize the U-statistic so that the private error term is smaller than the non-private error term $\mathrm{Var}(U_n)$?

Reweighting the data

Private U-statistic?

How can we privatize the U-statistic so that the private error term is smaller than the non-private error term $\mathrm{Var}(U_n)$?

Idea: Weighted U-statistic!

$$\hat{U}_n(X_1, \ldots, X_n) = \binom{n}{k}^{-1} \sum_{S \subseteq \binom{[n]}{k}} w(X_S) h(X_S).$$

Want: X_S is "atypical" $\iff$ $w(X_S)$ small

We will re-weight the data points X_i themselves and define $w(X_S)$ as $\min_{i \in S} w(X_i)$ Similar idea appeared also in Ullman and Sealfon (2019).

Reweighting the data

Private U-statistic?

How can we privatize the U-statistic so that the private error term is smaller than the non-private error term $\mathrm{Var}(U_n)$?

Idea: Weighted U-statistic!

$$\hat{U}_n(X_1, \ldots, X_n) = \binom{n}{k}^{-1} \sum_{S \subseteq \binom{[n]}{k}} w(X_S) h(X_S).$$

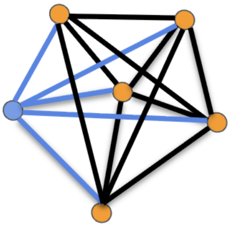
Want: X_S is "atypical" $\iff$ $w(X_S)$ small

We will re-weight the data points X_i themselves and define $w(X_S)$ as $\min_{i \in S} w(X_i)$ Similar idea appeared also in Ullman and Sealfon (2019).

Defining the weights $w(X_i)$

- For each index i, compute the U-statistic around X_i, or the local Hajek projection:

$$h_1^{(i)} = \binom{n-1}{k-1}^{-1} \sum_{S \subseteq \binom{[n]}{k}, X_i \in S} h(X_S).$$



Defining the weights $w(X_i)$

- For each index i, compute the U-statistic around X_i:

$$h_1^{(i)} = \binom{n-1}{k-1}^{-1} \sum_{S \subseteq \binom{[n]}{k}, X_i \in S} h(X_S).$$

- Let ξ and R be parameters such that

$$\Pr(\forall i, |h_1^{(i)} - U_n| \leq \xi) > 0.9, \quad \Pr(\forall S, |h(X_S) - U_n| \leq R) > 0.9.$$

- Define the weights

$$w(X_i; L) = \mathbb{1}\left(|h_1^{(i)} - U_n| \leq \xi + \frac{4kR}{n} \cdot L\right),$$

where L is the smallest positive integer such that at most L indices i have weight equal to 0.

Defining the weights $w(X_i)$

- For each index i, compute the U-statistic around X_i:

$$h_1^{(i)} = \binom{n-1}{k-1}^{-1} \sum_{S \subseteq \binom{[n]}{k}, X_i \in S} h(X_S).$$

- Let ξ and R be parameters such that

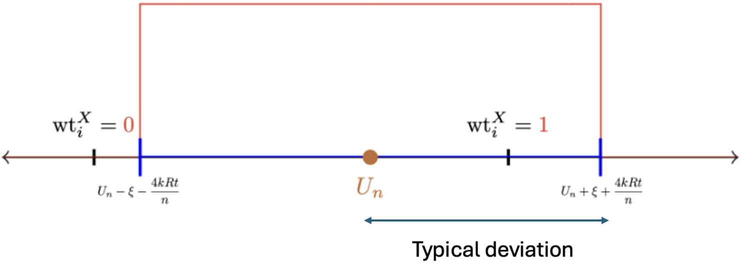
$$\Pr(\forall i, |h_1^{(i)} - U_n| \leq \xi) > 0.9, \quad \Pr(\forall S, |h(X_S) - U_n| \leq R) > 0.9.$$

- Define the weights

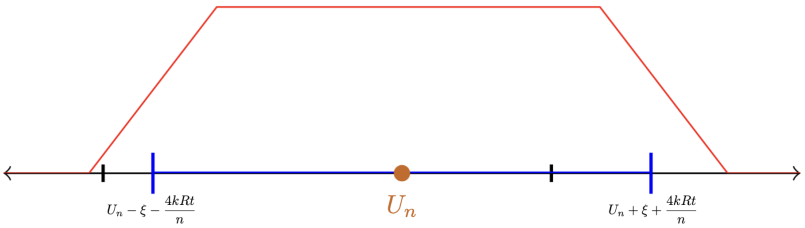
$$w(X_i; L) = \mathbb{1}\left(|h_1^{(i)} - U_n| \leq \xi + \frac{4kR}{n} \cdot L\right),$$

where L is the smallest positive integer such that at most L indices i have weight equal to 0.

Weight Function $w(X_i)$



Weight Function $w(X_i)$



Gives better guarantees than the previous weight function!

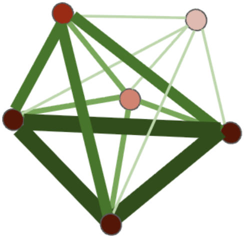
Private U-statistic

Once we have the weights $w(X_i)$ for all $i \in [n]$, define

$$w(X_S) = \min_{i \in S} w(X_i)$$

and the weighted U-statistic

$$U_w = \binom{n}{k}^{-1} \sum_S \left(w(X_S) h(X_S) + (1 - w(X_S)) U_n \right).$$



Weighted U-statistic

$$U_w := \binom{n}{k}^{-1} \sum_S \left(w(X_S) h(X_S) + (1 - w(X_S)) U_n \right).$$

- If all data points are "typical" then $L = 1$, all weights are equal to 1, and $U_w = U$.
- U_w curbs the sensitivity of bad points X_i by replacing the contributions of $h(X_S)$, $i \in S$, with U_n.

Punchline

We can employ the smoothed sensitivity mechanism instead of the Laplace mechanism. U_w has a smooth upper bound.

Private U-statistic: Algorithm

1. Compute the U-statistic U_n and the local U-statistics $h_1^{(i)}$
2. Assign a weight $w(X_i)$ to each data point depending on how close $h_1^{(i)}$ is to U_n and obtain L
3. Compute the weighted U-statistic U_w
4. Compute the β-smooth upper bound

$$SS(U_w(D)) = \max_{0\leq \ell \leq n} e^{-\beta L}\left(\frac{kL}{\beta}\left(\xi + \frac{kRL}{n}\right)\right),$$

where $\beta = \frac{\epsilon}{2\ln(2/\delta)}$.

5. Sample $Z \sim \mathrm{Lap}\,(1)$ and output the noisy U-statistic

$$\tilde{U}_n = U_w + \frac{2 \cdot SS(U_w)}{\epsilon} \cdot Z$$

Utility

Theorem 1

Our algorithm is $O(\epsilon)$-differentially private for any ξ. Moreover, suppose h is bounded with additive range C, and with probability at least 0.99, we have $\max_i |\hat{h}_1(i) - U_n| \le \xi$. There exists an algorithm such that, with probability at least $1 - \alpha$, we have

$$|\tilde{\theta} - \theta| = O\left(\sqrt{var(U_n)} + \frac{k\xi}{n\epsilon} + \left(\frac{k^2}{n^2\epsilon^2} + \frac{k^3}{n^3\epsilon^3}\right) C\right),$$

- Assume for simplicity $k = O(1)$
- Note that for non-degenerate U statistics, $\xi = \tilde{O}(1)$, $\sqrt{\mathrm{var}(U_n)} = O(1/\sqrt{n})$
- Note that for degenerate U statistics, $\xi = \tilde{O}(\sqrt{1/n})$ and $\sqrt{\mathrm{var}(U_n)} = O(1/n)$

Table

Algorithm	**Sub-Gaussian, non-degenerate**		**Bounded, degenerate**	
	Private error	Matches $O(\text{var}(U_n))$?	Private error	Matches $O(\text{var}(U_n))$?
Naive	$\tilde{O}\left(\frac{k\sqrt{\tau}}{n\epsilon}\right)$	No	$\tilde{O}\left(\frac{kC}{n\epsilon}\right)$	No
All-tuples	$\tilde{O}\left(\frac{k^{3/2}\sqrt{\tau}}{n\epsilon}\right)$	Yes	$\tilde{O}\left(\frac{kC}{n\epsilon}\right)$	No
Us	$\tilde{O}\left(\frac{k\sqrt{\tau}}{n\epsilon}\right)$	Yes	$\tilde{O}\left(\frac{k^{3/2}C}{n^{3/2}\epsilon}\right)$	Yes
Lower bound	$\Omega\left(\frac{k\sqrt{\tau}}{n\epsilon}\sqrt{\log\frac{n\epsilon}{k}}\right)$		$\Omega\left(\frac{k^{3/2}C}{n^{3/2}\epsilon}\right)$	

Table: We compare our application of off-the-shelf tools to our algorithm. We only provide the leading terms in the private error. The non-private lower bound on $\mathbb{E}(\hat{\theta} - \mathbb{E}h(X_1, \ldots, X_k))^2$ for all unbiased $\hat{\theta}$ is $\text{var}(U_n)$, which our private algorithms nearly match.

Roadmap

- Differential Privacy
- U statistics
- Random graphs

Recall latent geometric graphs

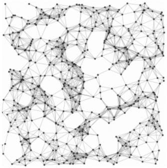


Figure: Latent geometric graph

- Consider n latent IID datapoints $X_i \in \mathbb{R}^d, i \in [n]$
- The edges are given by $A_{ij} \sim h(X_i, X_j)$
- We only observe $A_{ij}, 1 \leq i < j \leq n$
- Often one models sparsity by brining in a radius $h(x, y) = 1(\|x - y\|_2 \leq r_n)$ (Gilbert, 1961).
- Goal: We want to estimate the triangle density privately

Triangle density

- X_i is uniformly distributed on the surface of a 3-d sphere.
- The parameter $\theta_n = E[g(X_1, X_2, X_3)]$, where $g(x, y, z) = h(x, y)h(y, z)h(z, x)$
- Unbiased estimator:

$$U_n = \sum_{1 \le i < j < k \le n} \frac{g(X_i, X_j, X_k)}{\binom{n}{3}} = \sum_{1 \le i < j < k \le n} \frac{A_{ij} A_{jk} A_{ik}}{\binom{n}{3}}.$$

- Note that $\zeta_1 := \mathrm{var}(\mathbb{E}[g(X_1, X_2, X_3) | X_1]) = 0$ by symmetry.
- We can show that:

$$\mathrm{var}(U_n) \le \frac{r_n^4}{n^2}. \tag{1}$$

Applying off-the-shelf mean estimation

- Application of existing mean estimation algorithm Coinpress - w.h.p,

$$|\tilde{\theta}_{\text{coinpress}} - \theta| = \tilde{O}\left(\frac{r_n^2}{n} + \frac{1}{n\epsilon}\right),$$

  - As $r_n \to 0$, private error $\gg$ non-private error.

- Our algorithm achieves, w.h.p, as long as $r_n = \tilde{\Omega}(n^{-1/2})$,

$$|\tilde{\theta} - \theta| = \tilde{O}\left(\frac{r_n^2}{n} + \frac{1}{n^2\epsilon^2}\right),$$

  - Private error $\ll$ non-private error as long as $\epsilon = \Omega(r_n/\sqrt{n})$
  - This is possible because our algorithm captures the concentration of the local Hájek projections.

Summary

- There are a number of private algorithms for mean and covariance estimation
- We provide one for private estimation of an estimable parameter which can be written as $E[h(X_1, \ldots, X_k)]$. Our algorithm can also be applied to
  - Subsampled incomplete U statistics
  - Sub-gaussian $h(X_S)$
- This has applications in hypothesis testing, moment estimation in random geometric graphs
- It will be nice to extend to graphons, but it seems hard to obtain concentration of the local Hájek projections in graphons.

Folks

Kamalika Chaudhuri

Po-Ling Loh

Shourya Pandey

References I

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Gilbert, E. N. (1961). Random plane networks. *Journal of the Society for Industrial and Applied Mathematics*, 9:533–543.

Lee, A. J. (2019). *U-statistics: Theory and Practice*. Routledge.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, pages 75–84.

Ullman, J. and Sealfon, A. (2019). Efficiently estimating Erdos-Renyi graphs with node differential privacy. *Advances in Neural Information Processing Systems*, 32.