CS-5331 (4331) Assignment 1

Due on 02/17/2025, submit to Blackboard

1. **(30 points)** Figure 1 shows a medical record database released by a local hospital. In this database, "ZIP Code" and "Age" are the quasi-identifiers. The released database satisfies k-anonymity, distinct l_1-diversity, entropy l_2-diversity, t_1-closeness (t_1 is quantified using KL divergence), and t_2-closeness (t_2 is quantified using earth mover distance).

	ZIP Code	Age	Nationality	Disease
1	476**	2*	*	Heart Disease
2	476**	2*	*	Viral Infection
3	476**	2*	*	Cancer
4	476**	2*	*	Cancer
5	4790*	≥ 40	*	Viral Infection
6	4790*	≥ 40	*	Heart Disease
7	4790*	≥ 40	*	Viral Infection
8	4790*	≥ 40	*	Cancer
9	476**	3*	*	Cancer
10	476**	3*	*	Cancer
11	476**	3*	*	Viral Infection
12	476**	3*	*	Heart Disease

Figure 1: Medical record released by hospital.

   (a) (2 points) What is the value of k?

   (b) (2 points) What is the value of l_1?

   (c) (6 points) What is the value of l_2?

   (d) (8 points) What is the value of t_1?

   (e) (12 points) What is the value of t_2? Suppose $d_{ij} = 1, i, j \in$ {Heart Disease, Viral Infection, Cancer} when $i \neq j$. [Hint: You can use programming, like Python, or simplex algorithm to solve t_2.]

2. **(40 points)** Consider a normalized database collecting 2 attributes of 100 individuals, i.e., $X \in \mathcal{R}^{100\times 2}$ and $||x_i||_2 = 1, \forall i \in [1, n]$ (x_i denotes the sensitive attribute vector of the i-th individual, and all attribute vectors are normalized to have unit length). Let $q(X) = \frac{1}{100}X^T X$, which queries the empirical covariance matrix of the database.

   (a) (10 points) How to release the result of $q(x)$ satisfying 1-differential privacy using the Laplace mechanism? (Please give the l_1-sensitivity and the pdf of the calibrated distribution).

   (b) (10 points) How to release the result of $q(x)$ satisfying $(1, 10^{-5})$-differential privacy using the Gaussian mechanism? (Please give l_2-sensitivity and the pdf of the calibrated distribution).

   (c) (10 points) Generate 10,000 samples of i.i.d. noise for both distributions obtained in (a) and (b), plot and compare their histograms, and discuss which mechanism provides more utility.

   (d) (10 points) Use simple composition and advanced composition to evaluate the cumulative privacy loss (measured in terms of ϵ) when $q(x)$ is repeatedly calculated and then shared using the above Laplace and Gaussian mechanism. (Please plot the cumulative privacy loss versus increasing number of sharing. Set $\delta' = 10^{-6}$ in advanced composition.)

3. **(30 points)** Consider a database X collecting the salary of n individuals, i.e., $X \in \mathcal{R}^n$. The salary of each individual x_i is rounded to the nearest integer and $x_i \in [m] = \{1, 2, \ldots, m\}$. The database owner is interested in releasing the median value p of the salary by applying the Exponential Mechanism, and the score function is defined as

$$s(X, p) = -\left|\sum_{i \in [n]} Sign(x_i - p)\right|,$$

where $Sign(\cdot)$ denotes the sign function and $Sign(0) = 0$.

   (a) (5 points) Discuss the rationale of the defined score function. What is the optimal value of it?

   (b) (10 points) What is the sensitivity of the score function?

   (c) (15 points) Suppose that when the privacy budget is ϵ, the Exponential Mechanism $M_E(X, p, s)$ output differentially private median salary as $\widetilde{p}$. Let $index(\widetilde{p}, X \uparrow)$ be the index of $\widetilde{p}$ in $X \uparrow$ (i.e., the increasingly sorted version of X). Prove the following holds,

$$\Pr\left[\left|index(\widetilde{p}, X \uparrow) - \frac{n}{2}\right| \geq \frac{4(\ln(m/\alpha))}{\epsilon}\right] \leq \alpha.$$

4. **(40 points)**[1] A health insurance company wants to estimate of the fraction HIV+ in population of N patients. Let $X_i \in \{0, 1\}$ be the response of the ith patient (1 indicates HIV+ and 0 indicates HIV-). Suppose all patient will adopt random response to preserve their privacy, i.e., tell the truth X_i with probability $\frac{1}{2} + \gamma$ and tell a lie $1 - X_i$ with probability $\frac{1}{2} - \gamma$, where $\gamma \in (0, \frac{1}{2})$. Use Y_i to indicate the response of the ith patient, we have

$$Y_i = \begin{cases} X_i & \text{with probability } \frac{1}{2} + \gamma \\ 1 - X_i & \text{with probability } \frac{1}{2} - \gamma \end{cases}.$$

Denote the true HIV+ percentage as $p = \frac{1}{N}\sum_{i=1}^{N} X_i$. Answer the following.

   (a) (5 points) Random response achieves ϵ-DP. Show the value of ϵ.

   (b) (15 points) Justify that if the health insurance company directly uses $\tilde{p} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ to approximate p, it will lead to bias. Instead of using $\tilde{p}$ directly, help the company construct an unbiased estimator $\tilde{p_u}$, such that $\mathbb{E}[\tilde{p_u}] = p$.

   (c) (5 points) Show that $\mathrm{Var}[\tilde{p_u}] \leq \frac{1}{16\gamma^2 N}$, where $\tilde{p_u}$ is your constructed unbiased estimator in (b).

   (d) (5 points) Discuss why $\tilde{p_u}$ and $\tilde{p}$ have the exact same DP guarantee.

   (e) (10 points) Prove the following utility guarantee $|\tilde{p_u} - \tilde{p}| = O(\frac{1}{\gamma\sqrt{N}})$. [Hint: use Chebyshev's inequality.]

[1] This question is optional for undergraduate students enrolled in CS 4331 and can be attempted for extra credit. However, it is mandatory for graduate students enrolled in CS 5331. The total score for this assignment, including this question, is 140, but for graduate students, it will be normalized to 100.