COMP 430/530: Data Privacy and Security - Fall 2020
Homework Assignment # 3

General rules:

- You must show all your work and calculations. Correct answers without justification will receive little or no credit.
- In Questions 1-5, assume centralized DP. In Question 6, assume local DP.

Question 1 (Sensitivity). [15 pts] In each of the following scenarios, determine the *sensitivity* of the given query: $S(q)$. Give a concrete example in which this sensitivity is realized. Assume a neighboring dataset is obtained by adding or removing one individual.

(a) D is a trajectory dataset: $D = \{T_1, T_2, ..., T_{|D|}\}$, where T_i is one individual's trajectory consisting of $[1, T^{max}]$ location readings. Query q: How many location readings exist in Koc University campus?

(b) D is a set-valued dataset in which each individual makes at most N transactions. Buyable item universe is $I = \{i_1, i_2, ..., i_m\}$. Each item can be bought at most once per transaction. Query q: What is the support of an itemset?

(c) $D = (V, E)$ is an undirected, unweighted graph dataset in which each individual is a vertex and an edge represents that two individuals are connected. Let the degree of a vertex be $d(v)$, then the total degree of a graph is: $\sum_{v \in V} d(v)$. Query q: What is the total degree of a graph?

(d) D is a search log dataset in which each individual is associated with a series of keywords they searched for. An individual can make an arbitrary number of searches. Query q: How many users' search history includes the keyword "fenerbahce"?

(e) D is a tabular dataset in which each individual corresponds to one record. Query q: What is the number of t-close equivalence classes in the dataset?

Question 2 (Properties and implications of DP). [20 pts] Mathematically prove the following statements related to differential privacy.

Important: You must provide full mathematical proofs. You cannot say "this follows from the XYZ proof or XYZ property from the lecture".

Important: You cannot use sensitivity arguments in the proofs, since sensitivity-based methods are only a subset of methods that satisfy DP. Instead, use the probability-ratio definition of DP.

(a) Let $\mathcal{A}$ be a deterministic algorithm that does not output the same answer on all possible input datasets. Prove that $\mathcal{A}$ cannot satisfy ε-DP.

(b) Let $\mathcal{A}(D)$ be an ε-DP randomized algorithm that outputs a real-valued result, i.e, $\mathcal{A} : D \rightarrow \mathbb{R}$. Let f be a deterministic one-to-one mapping: $f : \mathbb{R} \rightarrow \mathbb{R}$. Prove that $f(\mathcal{A}(D))$ satisfies ε-DP.

(c) Let $\mathcal{A}_1(D)$ be an ε_1-DP algorithm such that $\mathcal{A}_1 : D \rightarrow \mathbb{R}^n$. Let $\mathcal{A}_2(D)$ be an ε_2-DP algorithm that is independent of $\mathcal{A}_1$ such that $\mathcal{A}_2 : D \rightarrow \mathbb{R}^n$. Prove that their sequential composition defined as $\mathcal{A}_{1,2}(D) = (\mathcal{A}_1(D), \mathcal{A}_2(D))$ satisfies $(\varepsilon_1 + \varepsilon_2)$-DP. **Hint:** Use independence of the algorithms.

(d) Let $\mathcal{A}$ be a randomized algorithm that satisfies ε-DP for neighboring datasets that differ in one individual. Prove that $\mathcal{A}$ satisfies $(k\varepsilon)$-DP for neighboring datasets that differ in k individuals.

Question 3 (DP or not?). [15 pts] Decide whether the following algorithms satisfy ε-DP or not. If they do, you must give a proof (similar to the examples in the lectures). If they do not, you must give a counter-example or explain why they do not satisfy DP. Correct answers without justification will receive little or no credit.

(a) Algorithm $\mathcal{A}$ takes as input a dataset D and counts the number of records in D. If the number of records is greater than e^{ε}, $\mathcal{A}$ prints out: "large". Otherwise, $\mathcal{A}$ prints out "small".

(b) Algorithm $\mathcal{A}$ takes as input a dataset D and an integer k. It returns TRUE if D is k-anonymous and FALSE otherwise.

(c) Algorithm $\mathcal{A}$ wants to answer if a certain individual's age is above 40. To do so, $\mathcal{A}$ takes as input dataset D and a person's name (eg: "John Doe") as input. It retrieves from D the count of individuals with that name and age > 40. Then, $\mathcal{A}$ adds Laplace noise with mean $= 0$ and scale $= \varepsilon$. If the noisy count is > 1, $\mathcal{A}$ returns TRUE. Otherwise, $\mathcal{A}$ returns FALSE.

Question 4 (DP Implementation). [15 pts] Recall the Adult dataset you cleaned and the histogram you built in Homework 1: "Create a histogram for understanding the education levels of individuals with high income (> 50K). Put the different education levels on the x axis. The y axis should be counts, i.e., the number of individuals with that education and income > 50K."

Here, we would like to build a differentially private version of this histogram. Use the clean version of Adult (remove missing values) and assume neighboring datasets are obtained via addition/removal of one individual: $D' = D \cup r$ or $D' = D \setminus r$, where r is one row.

(a) Describe a step-by-step algorithm based on the Laplace mechanism to build a differentially private histogram for the above task. Clearly state the steps of your algorithm, the scale of the Laplace noise you add, and whether your algorithm relies on any composition properties of DP.

(b) Implement your algorithm. Submit its source code.

(c) Run your private histogram algorithm with $\varepsilon = 0.01$ and plot its results side-by-side with the real histogram. What do you observe?

(d) Let H denote the actual histogram and $\hat{H}$ denote the private histogram. The average error in $\hat{H}$ can be measured bin-by-bin (bar-by-bar) as follows:

$$Err(\hat{H}, H) = \frac{\sum_b |\hat{H}[b] - H[b]|}{\text{number of bins}}$$

Generate private histograms for ε values of 0.01, 0.05, 0.1, 0.5, 1.0. Calculate and report their Err in a table or graph. What can you say about the relationship between ε and Err? Is this expected?

Question 5 (DP Implementation #2). [15 pts] Consider the same setting as Question 4, but this time, instead of creating a histogram, you would like to answer the question: What is the most common education level among those who earn income > 50K?

(a) Extend your algorithm from Question 4a to achieve this goal.

(b) Describe an algorithm based on the Exponential mechanism to achieve the same goal. Clearly state the steps of your algorithm, your choice of score function, and the sensitivity of your score function.

(c) Implement your algorithms from part a and b. Submit their source code.

(d) Run your Exponential-based algorithm and Laplace-based algorithm 100 times with ε values 0.0001, 0.001, 0.01, 0.1. How accurate are they with varying ε? Which one is preferable? You can measure their accuracy as the percentage of runs in which the correct outcome is returned.

Question 6 (Local DP). [20 pts] Consider the file *ages.txt* where each line corresponds to the age of one smartphone user. Assume ages are always between 1 and 100.

Say that each user's age is locally stored on their smartphone, and the data collector (server) wants to learn a histogram of users' ages. We will achieve this with local differential privacy.

(a) Implement the user-side data encoding and perturbation procedures using GRR.

(b) Implement the server-side estimation procedure for GRR to recover the noisy histogram.

(c) Using the error metric from Question 4d, what is the amount of error in the noisy histogram when the above procedure is simulated for $\varepsilon = 0.5, 1.0, 2.0, 4.0$? Are these results expected?

(d) For the same task, let us now assume we want to accomplish it with Simple RAPPOR instead of GRR. Implement the user-side data encoding and perturbation procedures using Simple RAPPOR.

(e) Implement the server-side estimation procedure for Simple RAPPOR to recover the noisy histogram.

(f) Using the error metric from Question 4d, what is the amount of error in the noisy histogram when the procedure with Simple RAPPOR is simulated for $\varepsilon = 0.5, 1.0, 2.0, 4.0$?

(g) Which protocol is preferable: GRR or Simple RAPPOR? Why?

(h) Describe how you would extend either of the two above approaches (GRR or Simple RAPPOR) to estimate the average age of the population.