CS848 Winter 2021 Assignment 1

[**TODO**]: [Student Name]

Due by Feb 1, 11pm

Please submit a pdf solution on Learn: "Submit" → "Dropbox folder" → "Assignment 1".

Q1. Laplace Mechanism

Consider a query that takes in a database D from domain $\mathcal{D}$ and outputs a vector of real number, $q : \mathcal{D} \rightarrow \mathbb{R}^d$. Laplace mechanism perturbs the query output with a vector of noise $\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_d \end{bmatrix}$, where $\eta_i \sim Lap(\lambda)$, i.e., $\Pr[\eta_i] \propto e^{-|\eta_i|/\lambda}$.

(a) Privacy Guarantee

[5 points] Show that the above Laplace mechanism achieves ϵ-differential privacy, if $\lambda = GS(q)/\epsilon$, where $GS(q) = \max_{neighors(D_1,D_2)} \|q(D_1) - q(D_2)\|_1$.

[**TODO**]: Add your proof here.

(b) Utility Guarantee

[5 points] If measuring the accuracy of this mechanism in terms of the sum of the expected squared errors, i.e.

$$\mathcal{E} = \sum_{i=1}^{d} \mathbb{E}[(o_i - c_i)^2],$$

where o_i is the ith entry of the noisy output, and c_i is the ith entry of the true answer. If $\lambda = GS(q)/\epsilon$, what is $\mathcal{E}$? What does this error depend on?

[**TODO**]: Add your analysis here.

Q2. Global Sensitivity Analysis

Consider a table of edges (e.g. Table 2) for an undirected graph of n nodes. Each row in the table represents an edge (v_i, v_j) in the graph, the node with smaller index stores in the first column (source) and the node with bigger index stores in the second column (dest).

source	dest
v_1	v_2
v_1	v_3
v_2	v_3
$\vdots$	$\vdots$
v_{n-1}	v_n

Table 1: Edge Table

(a) Triangle Counting Query

[5 points] Triangle counting query counts the number of triangles in the graph (e.g. there is a triangle formed by the first three edges in Table 2.) What is a global sensitivity of a triangle counting query (i.e., the maximum change to the triangle count if adding a row to an edge table/removing a row from an edge table)? Please explain how you derive this answer. (Hint: you can use a worst case example, or you can analyze this query as a sequence of transformations (e.g. self-joins of the edge table)).

[TODO]: Add your analysis here.

(b) Degree Distribution Query

[5 points] A degree of a node is the number of edges incident to this node. The degree distribution of a graph is the list of counts for each degree values from 0 to $n-1$, as the largest possible degree is $n-1$. For example, if we have a graph of 4 nodes $\{v_1, \ldots, v_4\}$, and there are 2 edges $(v_1, v_2), (v_1, v_3)$, then the degrees for the four nodes are $deg(v_1) = 2, deg(v_2) = 1, deg(v_3) = 1, deg(v_4) = 0$. The degree distribution over degree ranges from 0 to 3 is [1,2,1,0], as degree 0 appears once, degree 1 appears twice, degree 3 appear 1 once, etc.

What is a global sensitivity of a degree distribution query (i.e., the maximum change to the degree distribution query if adding a row to an edge table/removing a row from an edge table)? Please explain how you derive this answer. (Hint: you can use a worst case example, or you can analyze this query as a sequence of transformations (e.g. groupby the edge table)).

[TODO]: Add your analysis here.

Q3. Privacy for A Database of Multiple Tables

[Bonus 5 points] Consider the edge table from Q2. If we have another node table that records the properties of each node (e.g. age, sex, income etc.)

node id	age	sex	income
v_1	19	f	10k
v_2	25	m	8k
v_3	24	f	4k
$\vdots$	$\vdots$		
v_n	30	f	6k

Table 2: Node Table

What is the global sensitivity of the following query "what is the degree distribution of the nodes who have income greater than 5k?"

(Hint: Shall we consider adding/removing a row from the node table or adding/removing a row from the edge table?)

[TODO]: Add your analysis here.