CSC 591: Privacy (Fall 2019)
Home Assignment #2
Unity ID: rkolhe

Question 1)

a)

The sensitivity of this query function will be 1. Reason- While generating an adjacent database, we can either remove a patient having HIV or remove a patient without having HIV. Removing a patient without HIV would not change the number of HIV patients in the database, but removing a patient with HIV would decrease the HIV patient count by 1. Hence, the maximum change that can occur in the database after removing one record/sensitivity is 1

b)

To ensure 0.01-differential privacy, we can perturb the output by adding a noise once we add some noise to answer to the query. Hence, we need to ensure that the following equation is true:

$$D(x) = f(x) + noise\ .$$

Here, D(x) is the value in the dataset, GS is the global sensitivity function, where we are using a standard Laplace function to generate noise. Thus we can simplify the above equation as: $D(x) = f(x) + Laplace(\mu = 0,\ \lambda)$. Here, to maintain 0.01 DP, λ should be:

$$\lambda \geq sensitivty\ /\ \epsilon = 1/0.01 = 100$$

We can assume $\lambda = 100$. Hence, the final equation will be: Here, $Laplace(0,\ 100)$ can be expressed as

$$pdf(x) = \frac{1}{2\lambda} exp\left(-\frac{|x|}{\lambda}\right) = \frac{1}{200} exp\left(-\frac{|x|}{100}\right)$$

c)

In this question, our privacy budget is 0.01. Hence, to maintain privacy over 100 queries, we need to maintain $\epsilon/k = 0.01/100 = 10^{-4}$, privacy per query. Hence, ϵ should be **0.0001** for each query to maintain 0.01 DP over 100 queries.

Question 2)

a)

Sensitivity is defined as the maximum change that can occur in the dataset when we generate an adjacent dataset. Assume that the original dataset is D whereas the adjacent datatset is

D'.Hence, global sensitivity can be defined as:

$$GS_f = max(f(x) - f(x'))$$

Here, $f(x)$ is given as $mean(x)$. Hence, $f(x)$ will be the mean of the original dataset whereas $f(x')$will be the mean value of the adjacent dataset. To find the max difference bewteen the two values, we need to do the following:

1. Find the mean from D, $i.e\ f(x)$
2. Remove one random value from D to generate D'
3. Calculate the new mean $f(x')$

Now, $f(x)$ can be calculated as $\frac{\sum_{i=1}^{n} x_i}{n}$ assuming that there n values in the datatset. Now, we need to decide which value to be removed to get the max difference between the two datasets. We know that the range of values of the dataset is $[a, b]$. We must check first by removing the max value from the dataset and then by removing the min value from the dataset. Since the range of dataset is $[a, b]$, the max value is b and the min value is a. We need to consider the following two cases

Case I: Remove b

$$sensitivity_b = abs\left(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - b}{n-1}\right)$$

Case II: Remove a

$$sensitivity_a = abs\left(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - a}{n-1}\right)$$

Now, to select sensitivity, we will take the maximum of both the above cases:

$$Sensitivty = max\left[abs\left(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - b}{n-1}\right), abs\left(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - a}{n-1}\right)\right]$$

Assuming number of values in the database = n

b)

The query response will add noise once we calculate answer $f(D)$ to the query. To maintain $\epsilon - DP$, we can perturb the output by adding noise using Laplace distribution $(\mu = 0,\ \lambda)$. To ensure $eps - D$, the value of λ should be as follows

$$\lambda \geq |h - h'|/\epsilon \geq \frac{abs\left(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - b}{n-1}\right)}{\epsilon}$$

OR

$$\lambda \geq |h - h'|/\epsilon \geq \frac{abs\left(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - a}{n-1}\right)}{\epsilon}$$, based on the sensitivity chosen from above

$$i.e.\ Sensitivty = max\left[abs\left(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - b}{n-1}\right), abs\left(\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i - a}{n-1}\right)\right]$$

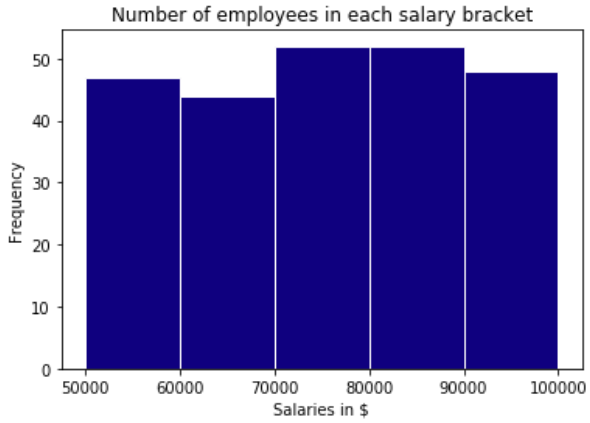
Hence, if we assume $\lambda = \frac{sensitivity}{\epsilon}$, our query response using Laplace distribution with parameter λ will be:

$f(D) + Laplace\left(\mu = 0, \lambda = \frac{sensitivity}{\epsilon}\right)$. We can substitue the value of λ in the pdf distribution. The Laplace distribution can be expressed as:

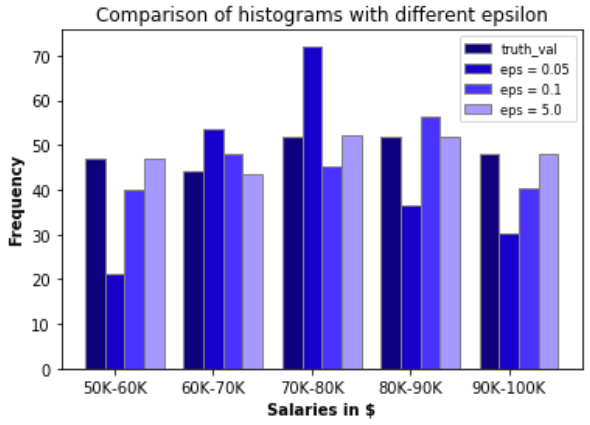
$$pdf(x) = \frac{1}{2\lambda} exp\left(-\frac{|x|}{\lambda}\right) = \frac{1}{2 * \frac{sensitivity)}{\epsilon}} exp\left(-\frac{|x|}{\frac{sensitivity}{\epsilon}}\right)$$

Question 3) (submitted code file)

a)



b)



c)

i)

When we increase the value of ϵ, we can observe that the magnitude of change of noise decreases based on the random value generated for smaller values of ϵ. When the value of ϵ is 0.05, the noise is high and when the value of ϵ increases to 0.1 and 5.0, the magnitude of change of noise is smaller as compared to the previous value of ϵ.

ii)

As the value of ϵ increases, the magnitude of noise decreases. Hence, as we increase ϵ, utility of histogram increases but, privacy is decreased.

Question 4)

If we express the 16-bit data using a 2-server PIR (Private Information Retrieval) protocol under O($n^{1/2}$) scheme, the 11-th position will be the highlighted cell, with rows $\Longrightarrow$ $i = 11/4 = 2$ and cols$\Longrightarrow j = 11\%4 = 3$

0	1	1	0
0	1	0	1
1	0	0	1
1	0	1	1

For each of the two servers, we will treat each row as a single dataset and find dot product of each row with a random string say Q_k and xor the result:

For server1- $S1$

Let $Q_1 = 1010$ $(random\ string)$. Now, take a dot product of Q_1 with each row and xor result. Therefore,

The dot products of Q_1 with each row will be as follows:

	j$\Longrightarrow$0	j$\Longrightarrow$1	j$\Longrightarrow$2	j$\Longrightarrow$3	XOR value of columns- Q_1
i$\Longrightarrow$0	0	0	1	0	1
i$\Longrightarrow$1	0	0	0	0	0
i$\Longrightarrow$2	1	0	0	0	1
i$\Longrightarrow$3	1	0	1	0	0

Hence, $Q_1 = < 1010 >$. We then, send Q_1 to the user. Hence, S_1 will send 1010

Now we XOR Q_1 with row with row j = 3, having the bit that is requested by the user. This will give us Q_2, which will be sent to Server 2- S_2

$Q_2 = 1010 \oplus 0001 = 1011$

Now, we take dot product of Q_2 with each row and xor the result. Therefore dot products of Q_2 with each row will be as follows:

	j⟹0	j⟹1	j⟹2	j⟹3	XOR value of columns- Q_2
i⟹0	0	0	1	0	1
i⟹1	0	0	0	1	1
i⟹2	1	0	0	1	0
i⟹3	1	0	1	1	1

Now, $Q_2 = 1101$. We now, send Q_2 to the user. Hence, Q_2 will send 1101

The user will receive 1010 from S_1 and 1101 from S_2. The highlighted element was present in $row\ i \implies 2$. Thus, the user will extract the bit at $i \implies 2$ from both the results. Therefore, it will extract 1 from S_1 and 0 from S_2. Now, the user will xor these values and retrieve the desired result.

$\implies 1 \oplus 0 = 1$

We can conclude that privacy is maintained since the user successfully received the correct information without broadcasting the desired row/column to the servers.