Homework 2

Nida Syed
nsyed@ncsu.edu

30 September 2019

1. Problem 1

   (a) Let D be the original dataset and D' be the modified dataset with one positive HIV record removed or modified.

   f is the query function, *"How many patients are HIV positive?"*

   Let h be the number of HIV positive patients in D, and h' be the number of HIV positive patients in D'.

   Therefore $f(D) = h$ and $f(D') = h'$.

   We know that sensitivity is the maximum difference over all pairs of datasets in D and D' differing in at most one element.

   Intuitively, the sensitivity of the query function is 1, since changing any one of the entries in the database causes the output of the function to change by either 0 or 1.

   Sensitivity for the database is $\left|h - h'\right| = 1$

   (b) The query response, should be the output with noise(generated from a Laplace function).
   This is $A(D) = f(D) + Lap(\frac{GS_f}{\varepsilon})$

   We know that, $\varepsilon = 0.01$.
   We computed GS_f, the global sensitivity to be $\left|h - h'\right| = 1$.

   Thus, we get $A(D) = f(D) + Lap(\frac{\left|h-h'\right|}{\varepsilon})$, where $\lambda = \frac{\left|h-h'\right|}{\varepsilon}$
   $= f(D) + Lap(\frac{1}{0.01})$

   Query response, $A(D) = f(D) + Lap(100)$, where $\lambda = 100$.

(c) Overall budget, $\varepsilon_0 = 0.01$
Number of queries, $k = 100$
ε for each query $= \frac{\varepsilon_0}{k}$

$\varepsilon = \frac{0.01}{100} = 0.0001$

Each query should be 0.0001 differentially private.

2. Problem 2

(a) We have a dataset $D = [a, b]$ with n salaries. Let us assume that it is sorted with $min(D) = a$ and $max(D) = b$.

Let us have a modified dataset D' with the entire sorted dataset in D with b removed, so it has $n - 1$ salaries. We have $min(D') = a$ and $max(D') = b'$ which is less than b.

Let S_n be the sum of all the salaries in $D = [a, b]$.

$mean(D) = \overline{D}$

$$= \frac{a + ... + b' + b}{n} = \frac{S_n}{n}$$

$mean(D') = \overline{D'}$

$$= \frac{a + ... + b'}{n - 1} = \frac{S_n - b}{n - 1}$$

Since f is the $mean$ function, $f(D) = \overline{D}$ and $f(D') = \overline{D'}$.

We know that sensitivity is the maximum difference over all pairs of datasets in D and D' differing in at most one element.

Sensitivity $= \left|f(D) - f(D')\right| = \left|mean(D) - mean(D')\right|$

Thus, $\overline{D} - \overline{D'}$

$$= \frac{a + ... + b' + b}{n} - \frac{a + ... + b'}{n - 1} = \frac{S_n}{n} - \frac{S_n - b}{n - 1}$$

Sensitivity is $\left|f(D) - f(D')\right|$

$$= \left|\frac{S_n}{n} - \frac{S_n - b}{n - 1}\right|$$

(b) The query response algorithm for the database D is,
$San(D) = f(D) + \xi$

$f(D)$ returns the *mean* of the salaries in D.
ξ is a Laplacian distribution with variance that depends on the sensitivity of function f and the privacy parameter ε.

$San(D) = mean(D) + Lap(\frac{GS_f}{\varepsilon})$

We computed GS_f, the global sensitivity to be $\left|f(D) - f(D')\right|$

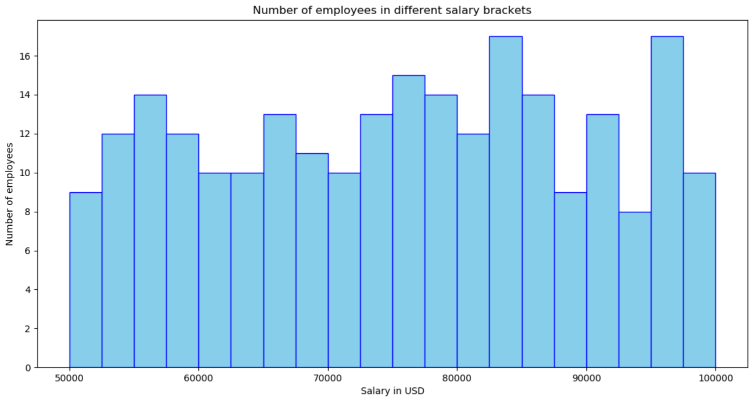
$$= \left| \frac{S_n}{n} - \frac{S_n - b}{n-1} \right|$$

where S_n is the sum of all salaries in D, b is the highest salary in D, and n is the total number of salaries in D.
Thus, we get the query response,

$$San(D) = mean(D) + Lap(\frac{\left|\frac{S_n}{n} - \frac{S_n - b}{n-1}\right|}{\varepsilon})$$

3. Problem 3

(a) Computed a histogram of the number of employees in different salary brackets.



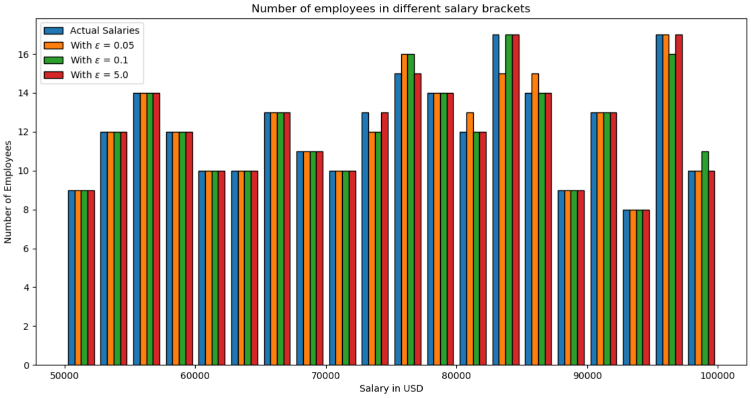
(b) The sensitivity for a histogram is 2.* We know that the scale of noise, $\lambda = Sensitivity/\varepsilon$.

When, $\varepsilon = 0.05$, $\lambda = \frac{2}{0.05} = 40$.

When, $\varepsilon = 0.1$ $\lambda = \frac{2}{0.1} = 20$.

When, $\varepsilon = 5.0$, $\lambda = \frac{2}{5.0} = 0.4$.

Computed ε-differentially private histograms for ε=0.05, 0.1 and 5.0.



(c) We see that with an increase in ε, the output perturbation due to noise ie, the distortion in the histogram decreases, and it resembles the original dataset. With an increase in ε, the privacy of the dataset decreases.

The utility of the histogram increases with an increase in ε. This is because the distortion decreases with an increase in ε, and it resembles the original dataset. In the histograms above, when $\varepsilon = 5.0$, it is the same as the original histogram, and it has no perturbation, and the utility is maximum

4. Problem 4

We need to find the retrieve the bit in the 11th position in the following data of length $n = 16$:

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

The index of the bit to be retrieved, $k = 11$

Converting this linear array into a 2D array, we get, $\sqrt{n} = 4$:

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

We have a $4x4$ matrix, hence $11/4 = 2 = i$ and $11\%4 = 3 = j$.

The bit to be retrieved lies in the *ith* row and *jth* column.

Hence we need to retrieve the bit that lies in the row with index 2 and column with index 3.

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

There are two servers, $S1$ and $S2$, who have this data.

$S1$:

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

$S2$:

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

With $S1$ data, let us perform row-wise computation of dot product and XORing the result with $Q1$, a random string of length $\sqrt{n} = 4$.

$Q1 = 1010$

$0.1 \oplus 1.0 \oplus 1.1 \oplus 0.0 = 1$
$0.1 \oplus 1.0 \oplus 0.1 \oplus 1.0 = 0$
$1.1 \oplus 0.0 \oplus 0.1 \oplus 1.0 = 1$
$1.1 \oplus 0.0 \oplus 1.1 \oplus 1.0 = 0$

We have $S1.Q1$:

1
0
1
0

With $S2$ data, we perform row-wise computation of dot product and XORing the result with $Q2$, where $Q2 = Q1 \oplus j$.

As $j = 3$, it is represented as 0001, with the $3rd$ bit as 1.

$Q2 = 1010 \oplus 0001 = 1011$

$0.1 \oplus 1.0 \oplus 1.1 \oplus 0.1 = 1$
$0.1 \oplus 1.0 \oplus 0.1 \oplus 1.1 = 1$
$1.1 \oplus 0.0 \oplus 0.1 \oplus 1.1 = 0$
$1.1 \oplus 0.0 \oplus 1.1 \oplus 1.1 = 1$

We have $S2.Q2$:

1
1
0
1

XORing the ith row of $S1.Q1$ and $S2.Q2$:

1	1
0	1
1	0
0	1

$1 \oplus 0 = 1$

1 is now sent to the querier without the servers knowing what the position of the bit was.

*__References:__

http://dimacs.rutgers.edu/~graham/pubs/slides/privdb-tutorial.pdf