Alp Kural – 71959
Computer Engineering

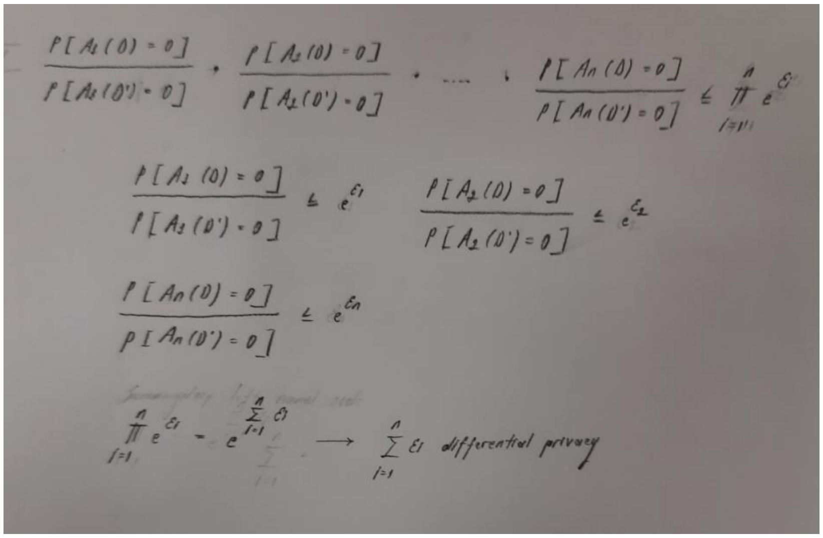
Comp 430 – Homework 02 Project Report

PART 1 / PART A

Disclosure of one of the resident's sensitive information by his/her consent does not violate differential privacy. Since, his/her information is revealed by his/her consent and with the known fact that the probability of output of a query with a database and the same database subtracted the entry related to the citizen whose sensitive information is disclosed (neighboring database) will be smaller than e^{ε} which guarantees differential privacy with the privacy parameter ε.

PART 1 / PART B

In the context of differential privacy, when we say that algorithms are independent, we are referring to the fact that the outcomes of each algorithm are not influenced by the presence or absence of the others. The inequalities' right-hand side and left-hand side must be multiplied because we are examining a probability which has multiple independent events. By multiplying differential privacy inequalities' right-hand side and left-hand side corresponding to each individual algorithm, it can be proved that the sequential composition of algorithms satisfies DP with the privacy parameter being sum of each privacy parameter corresponding to respective algorithm.



$$\frac{P[A_1(D)=0]}{P[A_1(D')=0]} \cdot \frac{P[A_2(D)=0]}{P[A_2(D')=0]} \cdot \dots \cdot \frac{P[A_n(D)=0]}{P[A_n(D')=0]} \leq \prod_{i=1}^{n} e^{\varepsilon_i}$$

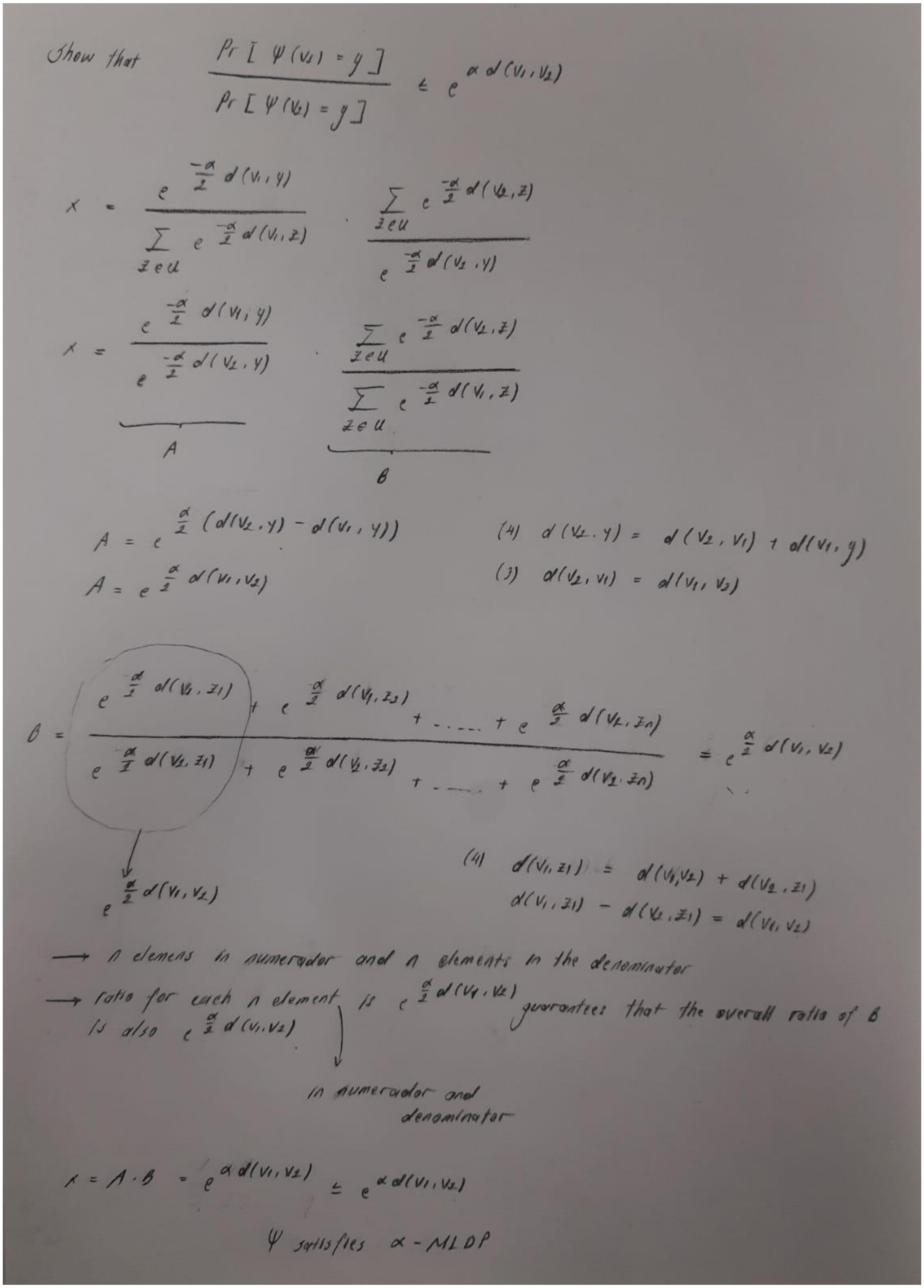
$$\frac{P[A_1(D)=0]}{P[A_1(D')=0]} \leq e^{\varepsilon_1} \qquad \frac{P[A_2(D)=0]}{P[A_2(D')=0]} \leq e^{\varepsilon_2}$$

$$\frac{P[A_n(D)=0]}{P[A_n(D')=0]} \leq e^{\varepsilon_n}$$

$$\prod_{i=1}^{n} e^{\varepsilon_i} = e^{\sum_{i=1}^{n} \varepsilon_i} \longrightarrow \sum_{i=1}^{n} \varepsilon_i \text{ differential privacy}$$

Alp Kural – 71959
Computer Engineering

PART 2



Show that $\frac{Pr[\psi(v_1) = y]}{Pr[\psi(v_2) = y]} \leq e^{\alpha d(v_1, v_2)}$

$$x = \frac{e^{\frac{-\alpha}{2} d(v_1, y)}}{\sum_{z \in U} e^{\frac{-\alpha}{2} d(v_1, z)}} \cdot \frac{\sum_{z \in U} e^{\frac{-\alpha}{2} d(v_2, z)}}{e^{\frac{-\alpha}{2} d(v_2, y)}}$$

$$x = \underbrace{\frac{e^{\frac{-\alpha}{2} d(v_1, y)}}{e^{\frac{-\alpha}{2} d(v_2, y)}}}_{A} \cdot \underbrace{\frac{\sum_{z \in U} e^{\frac{-\alpha}{2} d(v_2, z)}}{\sum_{z \in U} e^{\frac{-\alpha}{2} d(v_1, z)}}}_{B}$$

$A = e^{\frac{\alpha}{2} (d(v_2, y) - d(v_1, y))}$ (4) $d(v_2, y) = d(v_2, v_1) + d(v_1, y)$

$A = e^{\frac{\alpha}{2} d(v_1, v_2)}$ (3) $d(v_2, v_1) = d(v_1, v_2)$

$$B = \frac{e^{\frac{\alpha}{2} d(v_1, z_1)} + e^{\frac{\alpha}{2} d(v_1, z_2)} + \ldots + e^{\frac{\alpha}{2} d(v_1, z_n)}}{e^{\frac{\alpha}{2} d(v_2, z_1)} + e^{\frac{\alpha}{2} d(v_2, z_2)} + \ldots + e^{\frac{\alpha}{2} d(v_2, z_n)}} = e^{\frac{\alpha}{2} d(v_1, v_2)}$$

$\downarrow$

$e^{\frac{\alpha}{2} d(v_1, v_2)}$

(4) $d(v_1, z_1) = d(v_1, v_2) + d(v_2, z_1)$

$d(v_1, z_1) - d(v_2, z_1) = d(v_1, v_2)$

→ n elemens in numerador and n elements in the denominator

→ ratio for each n element is $e^{\frac{\alpha}{2} d(v_1, v_2)}$ guarantees that the overall ratio of B is also $e^{\frac{\alpha}{2} d(v_1, v_2)}$

in numerador and denominator

$x = A \cdot B = e^{\alpha d(v_1, v_2)} \leq e^{\alpha d(v_1, v_2)}$

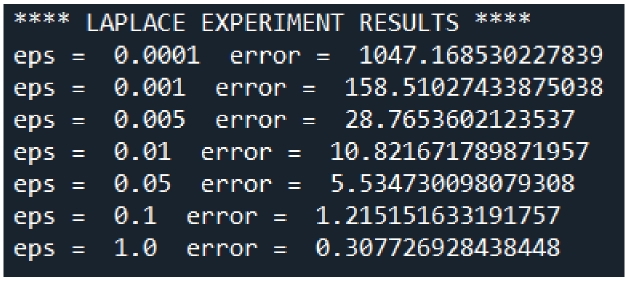
ψ satisfies α - MLDP

Alp Kural – 71959
Computer Engineering

PART 3 / TASK 01

Laplace Experiment Results

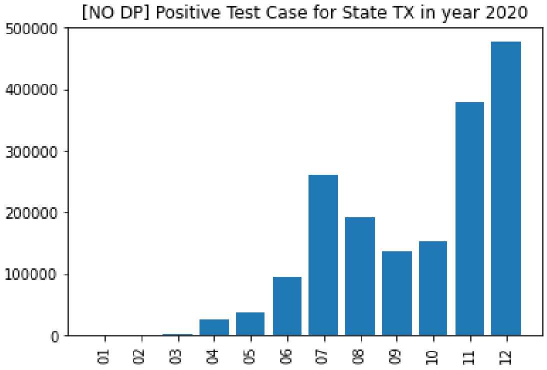
As the level of epsilon increases, signifying a more relaxed form of differential privacy, the perturbation of the differentially private histogram from the non-differentially private histogram also increases. Maintaining a higher degree of privacy (requiring low levels of epsilon) necessitates the introduction of increased noise, consequently resulting in higher error rates.



```
**** LAPLACE EXPERIMENT RESULTS ****
eps =  0.0001  error =  1047.168530227839
eps =  0.001  error =  158.51027433875038
eps =  0.005  error =  28.7653602123537
eps =  0.01  error =  10.821671789871957
eps =  0.05  error =  5.534730098079308
eps =  0.1  error =  1.215151633191757
eps =  1.0  error =  0.307726928438448
```

Epsilon	Error
0.0001	1047.17
0.001	158.51
0.005	28.77
0.01	10.82
0.05	5.53
0.1	1.22
1.0	0.31

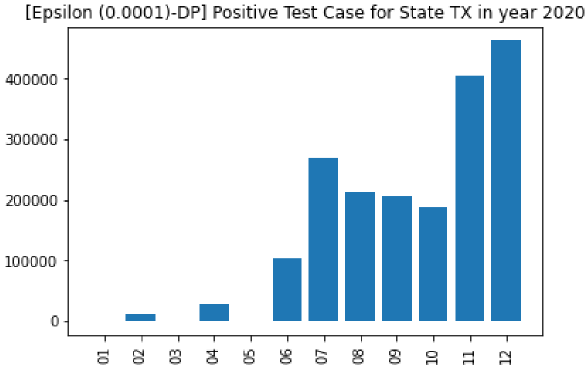
Non-Differentially Private Histogram

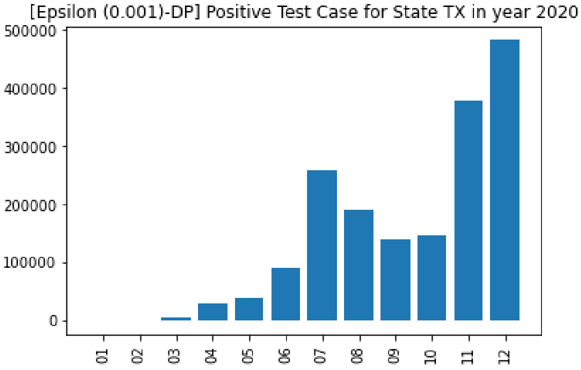


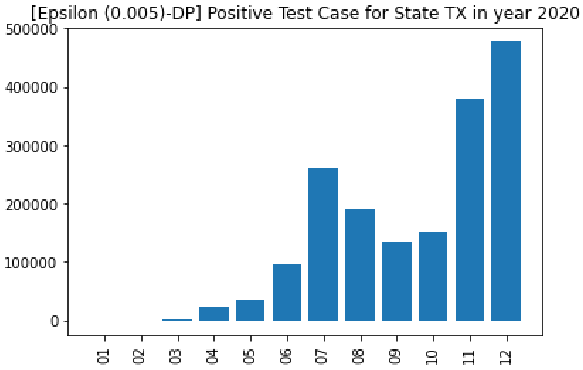
Differentially Private Histograms with the varying epsilon values listed as:

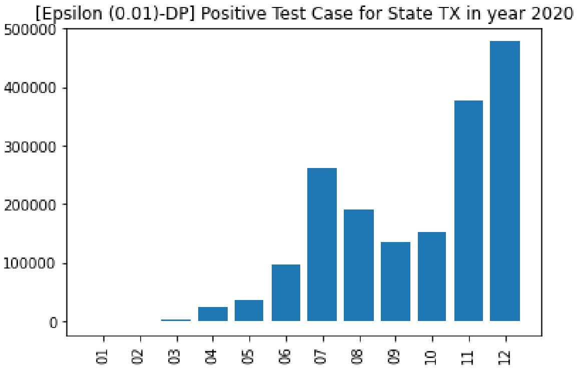
{0.0001, 0.001, 0. 005, 0.01, 0.05, 0.1, 1.0} (Graphs sampled from ten executions for each epsilon value.)

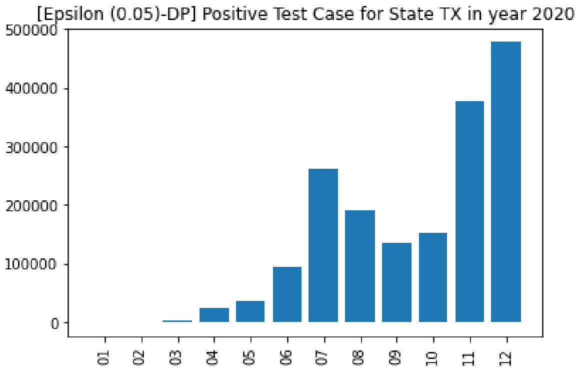
Corresponding epsilon value of each histogram is stated in the title of the plot.

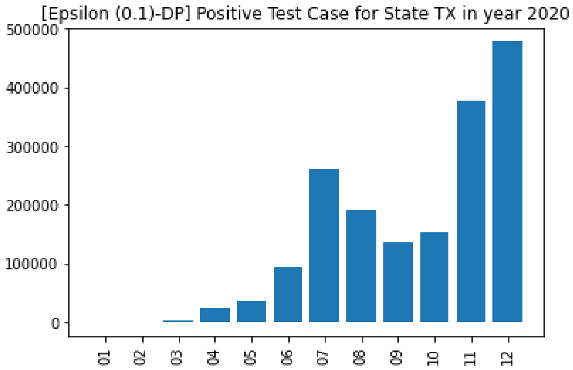


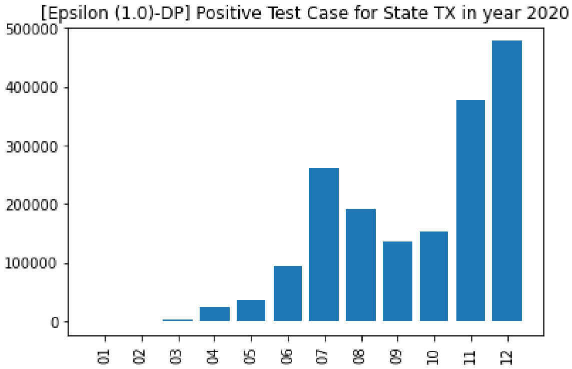



[Epsilon (0.005)-DP] Positive Test Case for State TX in year 2020

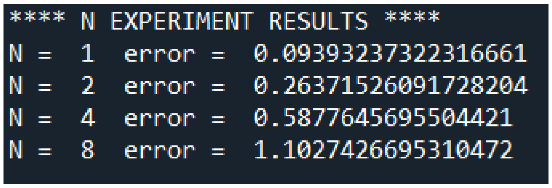

[Epsilon (0.01)-DP] Positive Test Case for State TX in year 2020


[Epsilon (0.05)-DP] Positive Test Case for State TX in year 2020


[Epsilon (0.1)-DP] Positive Test Case for State TX in year 2020


[Epsilon (1.0)-DP] Positive Test Case for State TX in year 2020

N Experiment Results



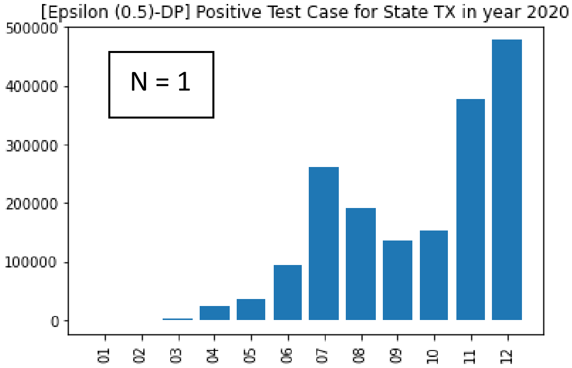
```
**** N EXPERIMENT RESULTS ****
N =  1  error =  0.09393237322316661
N =  2  error =  0.26371526091728204
N =  4  error =  0.5877645695504421
N =  8  error =  1.1027426695310472
```

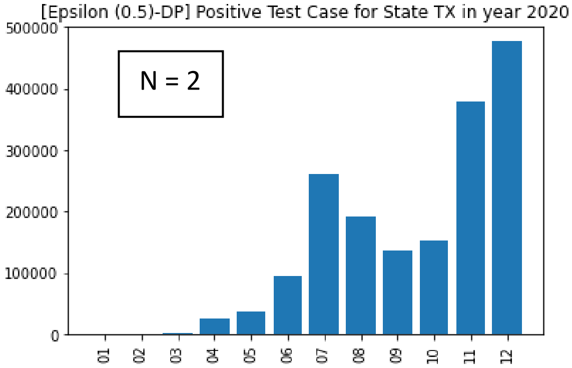
Sensitivity	Error
1	0.09
2	0.26
4	0.59
8	1.10

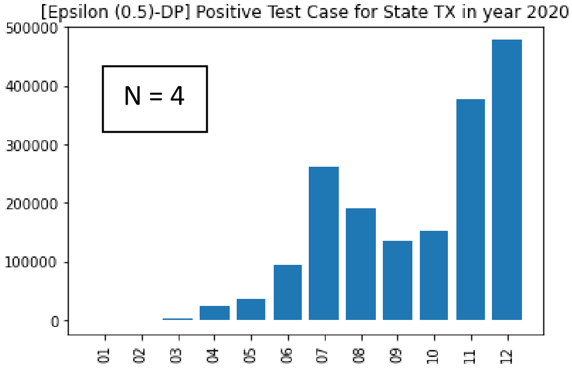
N value represents the sensitivity of information at each bin and as sensitivity increases the need for adding higher noise increase therefore the accumulated error becomes larges as N increases.

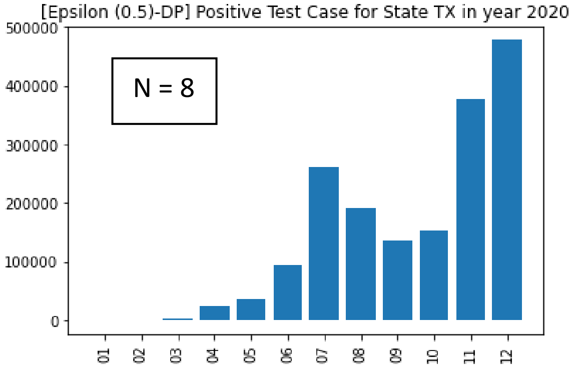
Differentially Private Histograms with the varying N values listed as:

{1, 2, 4, 8} (Graphs sampled from ten executions for each N value.)





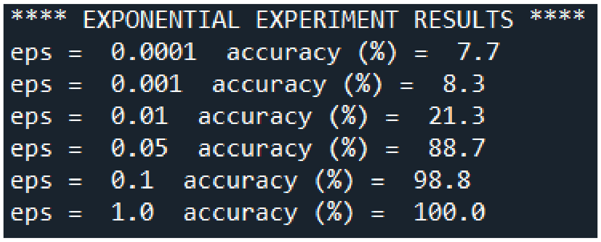




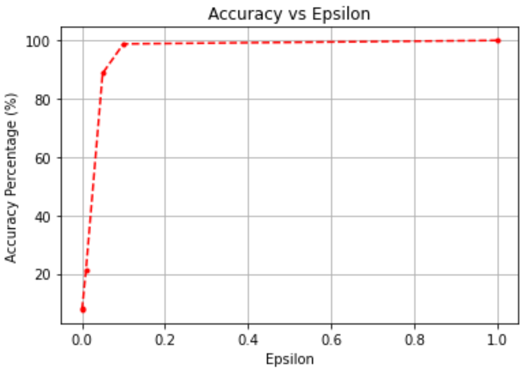
PART 3 / TASK 02

Thousand execution is conducted for each epsilon value.

Exponential Experiement Results



```
**** EXPONENTIAL EXPERIMENT RESULTS ****
eps =  0.0001  accuracy (%) =  7.7
eps =  0.001  accuracy (%) =  8.3
eps =  0.01  accuracy (%) =  21.3
eps =  0.05  accuracy (%) =  88.7
eps =  0.1  accuracy (%) =  98.8
eps =  1.0  accuracy (%) =  100.0
```



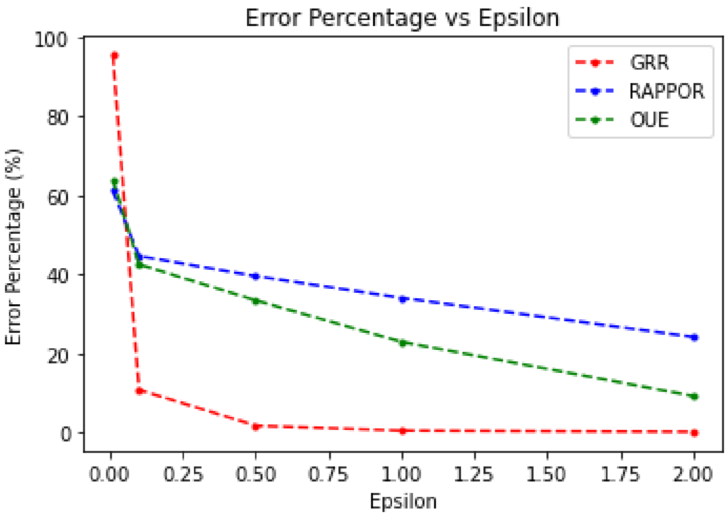
As the level of epsilon increases, signifying a more relaxed form of differential privacy, the probability to return the correct answer increases. To maintain a higher degree of privacy (requiring low levels of epsilon) accuracy to return the correct result becomes lower.

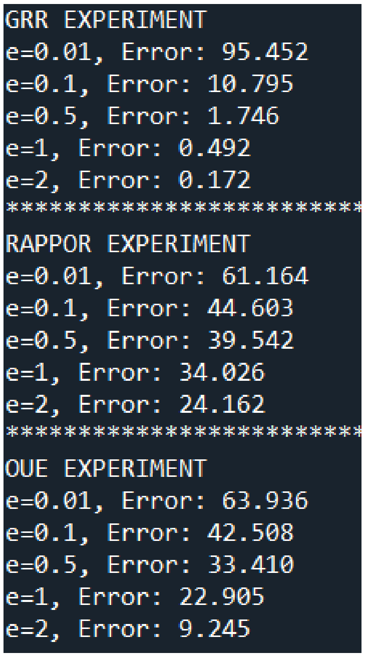
PART 4 / Experimental Analysis

Error Analysis

For each protocol (GRR, RAPPOR and OUE), as the demanded privacy from the protocol becomes more relaxed, the error percentage between the original data collected from users and the estimation of the data collector server of the perturbed data of the user becomes lower and lower.

It can be observed from the experimental analysis that from the perspective of accuracy GRR outperforms RAPPOR and OUE in most of the privacy conditions. Besides, from the experiment it can be stated that after GRR, OUE comes second in terms of accuracy. On the other hand, GRR demonstrates worst accuracy in the high privacy conditions like epsilon being 0.01. For an application, the protocol that will be used must be chosen with respect to highest accuracy in the demanded privacy condition.



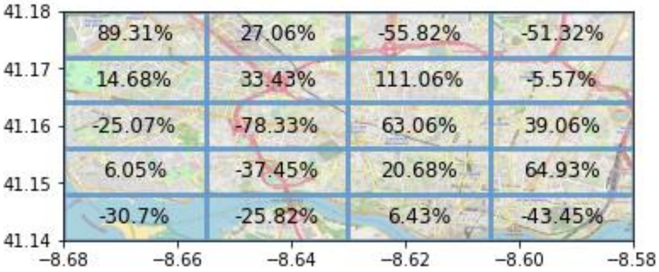


```
GRR EXPERIMENT
e=0.01, Error: 95.452
e=0.1, Error: 10.795
e=0.5, Error: 1.746
e=1, Error: 0.492
e=2, Error: 0.172
************************
RAPPOR EXPERIMENT
e=0.01, Error: 61.164
e=0.1, Error: 44.603
e=0.5, Error: 39.542
e=1, Error: 34.026
e=2, Error: 24.162
************************
OUE EXPERIMENT
e=0.01, Error: 63.936
e=0.1, Error: 42.508
e=0.5, Error: 33.410
e=1, Error: 22.905
e=2, Error: 9.245
```

Visual Analysis

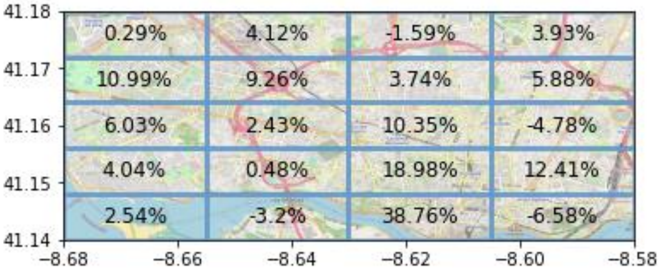
I picked the OUE protocol for the visual analysis.

Epsilon = 0.01



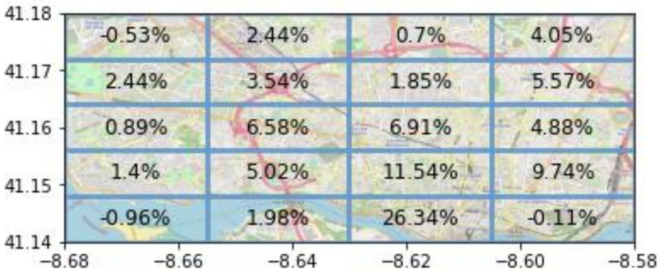
89.31%	27.06%	-55.82%	-51.32%
14.68%	33.43%	111.06%	-5.57%
-25.07%	-78.33%	63.06%	39.06%
6.05%	-37.45%	20.68%	64.93%
-30.7%	-25.82%	6.43%	-43.45%

Epsilon = 0.1

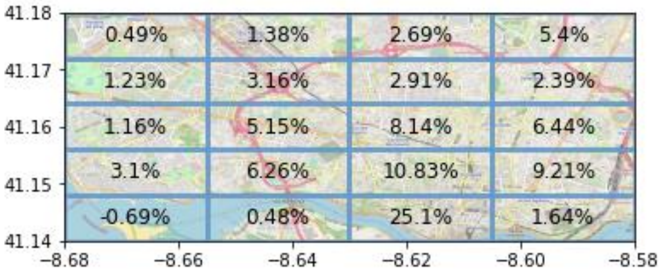


0.29%	4.12%	-1.59%	3.93%
10.99%	9.26%	3.74%	5.88%
6.03%	2.43%	10.35%	-4.78%
4.04%	0.48%	18.98%	12.41%
2.54%	-3.2%	38.76%	-6.58%

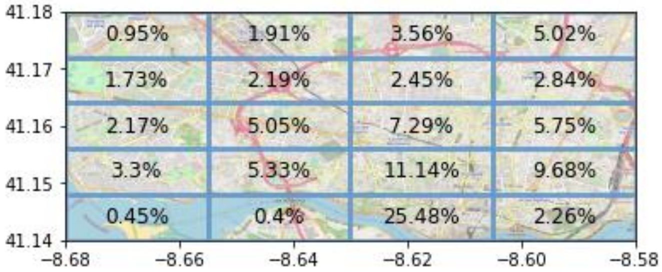
Epsilon = 0.5



Epsilon = 1



Epsilon = 2



From the observations of the plots, it can be stated that the distribution of the taxis in Porto becomes meaningful in the high epsilon values which indicates more relaxed privacy conditions due to the fact that accuracy increases with the relaxing privacy parameter. Besides, it can be observed that in the high privacy conditions the estimator estimates incorrectly negative percentages for some cells in the map.