COMP 530 Data Privacy and Security
Homework 2 Report

Due November 28th

Ismayil Ismayilov

Part 1: Privacy Proofs

Proof. As per the definition of α-$MLDP$ let us consider the ratio $\frac{Pr[\Psi(v_1)=y]}{Pr[\Psi(v_2)=y]}$ for input values $v_1, v_2 \in U$, output value $y \in U$ and whereby Ψ is the perturbation algorithm given in the problem statement.

$$
\begin{aligned}
\frac{Pr\left[\Psi(v_1)=y\right]}{Pr\left[\Psi(v_2)=y\right]} &= \frac{\frac{e^{\frac{-\alpha\cdot d(v_1,y)}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}}}{\frac{e^{\frac{-\alpha\cdot d(v_2,y)}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_2,z)}{2}}}} && \text{Plugging in values by the definition of } \Psi \\
&= \frac{e^{\frac{-\alpha\cdot d(v_1,y)}{2}}}{e^{\frac{-\alpha\cdot d(v_2,y)}{2}}} \cdot \frac{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_2,z)}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}} && \text{Rearranging terms} \\
&= X \cdot Y
\end{aligned}
$$

In the above derivation, we denote the term $\frac{e^{\frac{-\alpha\cdot d(v_1,y)}{2}}}{e^{\frac{-\alpha\cdot d(v_2,y)}{2}}}$ by X and the term $\frac{\sum_{z\in U} e^{\frac{-\alpha\cdot d(v_2,z)}{2}}}{\sum_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}}$ by Y and analyze them separately. Let us begin with X.

$$
\begin{aligned}
X &= \frac{e^{\frac{-\alpha\cdot d(v_1,y)}{2}}}{e^{\frac{-\alpha\cdot d(v_2,y)}{2}}} \\
&= e^{\frac{\alpha(d(v_2,y)-d(v_1,y))}{2}} \\
&\leq e^{\frac{\alpha(d(v_2,v_1)+d(v_1,y)-d(v_1,y))}{2}} && \text{Since } d(v_2,y) \leq d(v_2,v_1)+d(v_1,y) \text{ by the triangle inequality} \\
&= e^{\frac{\alpha\cdot d(v_2,v_1)}{2}} = e^{\frac{\alpha\cdot d(v_1,v_2)}{2}} && \text{Since } d(v_2,v_1) = d(v_1,v_2) \text{ by the symmetry property}
\end{aligned}
$$

We have shown that , $X \leq e^{\frac{\alpha\cdot d(v_1,v_2)}{2}}$. Let us now consider Y.

$$
\begin{aligned}
Y &= \frac{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_2,z)}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}} \\
&= \frac{\sum\limits_{z\in U} e^{\frac{-\alpha(d(v_2,z)+d(v_1,v_2)-d(v_1,v_2))}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}} && \text{Add and subtract } d(v_1,v_2) \text{ in the topmost numerator} \\
&\leq \frac{\sum\limits_{z\in U} e^{\frac{-\alpha(d(v_2,z)+d(v_1,z)+d(z,v_2)-d(v_1,v_2))}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}} && \text{Since } d(v_1,v_2) \leq d(v_1,z)+d(z,v_2) \text{ by the triangle inequality} \\
&= \frac{\sum\limits_{z\in U} e^{\frac{\alpha(d(v_1,v_2)-2d(v_2,z)-d(v_1,z))}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}}
\end{aligned}
$$

$$\begin{aligned}
&\leq \frac{\sum\limits_{z\in U} e^{\frac{\alpha(d(v_1,v_2)-d(v_1,z))}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}} = \frac{\sum\limits_{z\in U} e^{\frac{\alpha\cdot d(v_1,v_2)}{2}}\cdot e^{\frac{-\alpha\cdot d(v_1,z)}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}} \\
&= \frac{e^{\frac{\alpha\cdot d(v_1,v_2)}{2}}\cdot \sum\limits_{z\in U}\cdot e^{\frac{-\alpha\cdot d(v_1,z)}{2}}}{\sum\limits_{z\in U} e^{\frac{-\alpha\cdot d(v_1,z)}{2}}} && e^{\frac{\alpha\cdot d(v_1,v_2)}{2}} \text{ is constant in the summation} \\
&= e^{\frac{\alpha\cdot d(v_1,v_2)}{2}}
\end{aligned}$$

As for X, we have shown that $Y \leq e^{\frac{\alpha\cdot d(v_1,v_2)}{2}}$. Thefore, it must be that $\frac{Pr[\Psi(v_1)=y]}{Pr[\Psi(v_2)=y]} = X\cdot Y \leq e^{\frac{\alpha\cdot d(v_1,v_2)}{2}}\cdot e^{\frac{\alpha\cdot d(v_1,v_2)}{2}} = e^{\alpha\cdot d(v_1,v_2)}$. The proof is complete. □

Part 2: DP Implementation

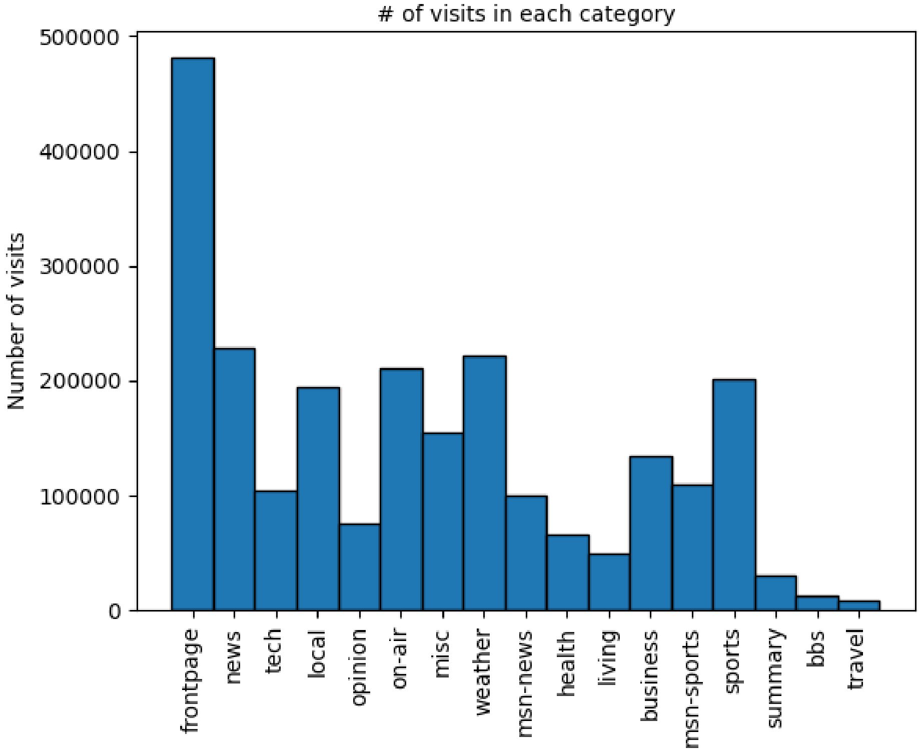


Figure 1: Non-private histogram

Err vs n

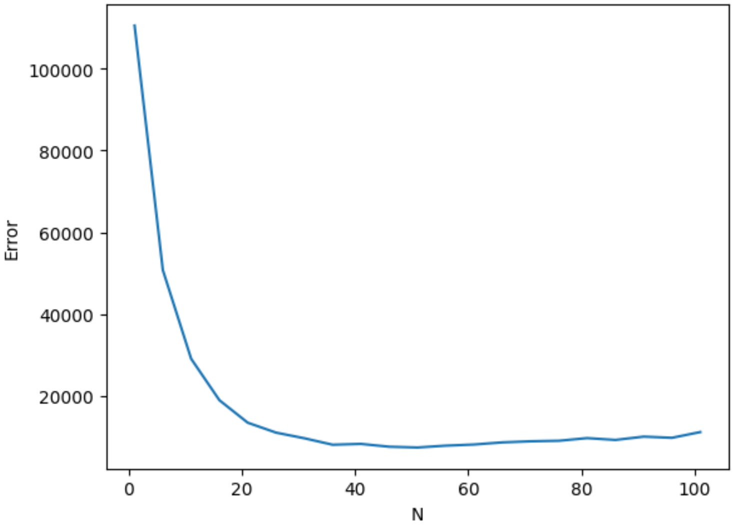


Figure 2: Error vs N

Truncating sequences is used as a means of controlling sensitivity. In this context, a small value of n drastically reduces sensitivity. The downside, however, is that a lot of valuable data points may be discarded which would lead to high error rates. This is confirmed by the figure, as smaller values of n result in high error rates which eventually level off as n increases. A large value of n on the other hand makes us of many data points but results in much higher sensitivity. Eventually as the sensitivity rises, the error rates start rising as well. Once more, this observation is confirmed by the figure; as n rises, the error rates fall to their global minimum but eventually start rising up again. Therefore, the fact that the lowest error is achieved for $n = 56$ makes sense intuitively. $n = 56$ is roughly at the midpoint of the range of n values; at the midpoint between low data utilization, low sensitivity and high data utilization, high sensitivity.

Err vs ϵ

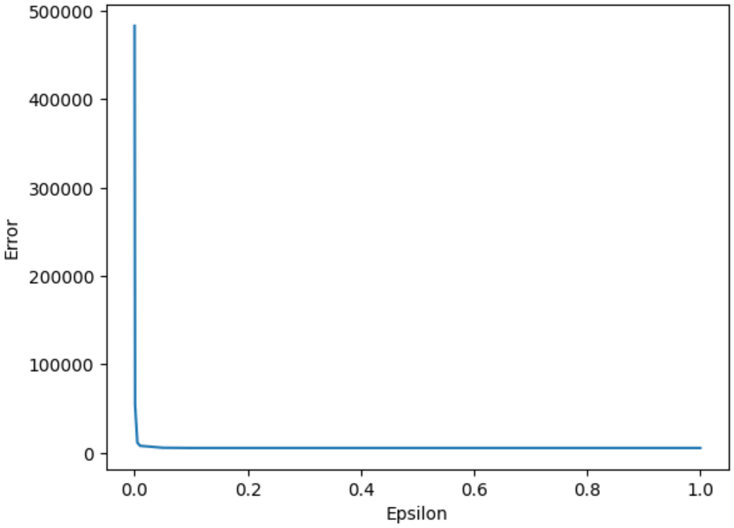


Figure 3: Error vs ϵ

As the graph, shows for $\epsilon = 0.0001$ and $\epsilon = 0.001$, the error is approximately 50.000 which can be explained by the overly strict privacy budget. As ϵ decreases, the error rates sharply drop down hovering around the minimum of 5500 for ϵ values $0.05, 0.1, 1.0$. In light of this, we can say that $\epsilon = 0.05$ is the best value for ϵ in this context as it the strictest value that results in the minimum error.

Accuracy vs ϵ

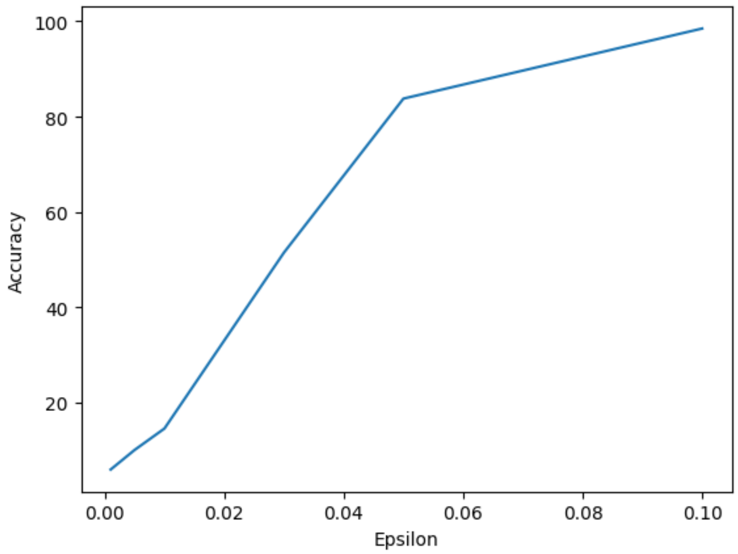
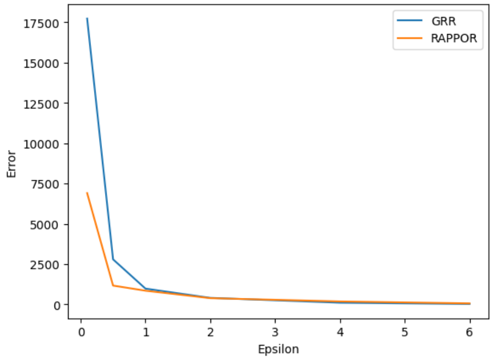


Figure 4: Accuracy vs ϵ

As the figure shows, small values of ϵ result in sub-par accuracy. As the privacy budget is relaxed, the accuracy steadily increases. The maximum accuracy of 99.2% percent is achieved for $\epsilon = 0.1$. In this context, we can say that $\epsilon = 0.05$ is the best value for the privacy budget as it is relatively strict and achieves an adequate accuracy of 83.2%.

Part 3: LDP Implementation



ϵ	GRR	RAPPOR
0.1	17734.19	6904.98
0.5	2792.50	1165.39
1.0	976.99	847.56
2.0	402.19	384.45
4.0	103.04	181.07
6.0	24.38	60.82

As figure and table above indicate, GRR results in very high error when the privacy budget is strict but eventually levels off as the privacy budget is relaxed. RAPPOR however achieves much lower error for strict privacy budgets but is eventually overtaken by GRR as the privacy budget is relaxed. It is a reasonable to conclude that if the privacy budget to be satisfied is strict, RAPPOR should be preferred. If, however, the privacy budget is lenient, GRR achieves lower error and should probably be used instead.

List of dependencies is provided in the *requirements.txt* file