Degree-Preserving Local Differential Privacy for Analyzing Coreness of Decentralized Social Networks

2023-2nd Semester

Undergraduate Thesis

Student ID.	20181127
Name	Jiwon Sung
Department	EE
Project Advisor	Sung Whan Yoon

FIRST IN CHANGE

I. Introduction

1) Topic and Purpose of Research

Differential privacy is a method that helps preserve the privacy of users, more specifically, allowing the data to be analyzed without leaking users' sensitive information. Privacy is guaranteed by having the central server, which has access to all the users' data, add noise to the data when releasing the data for analytics. However, when differential privacy is implemented, the users may not trust the central server and may be reluctant to send their personal data. Local differential privacy (LDP) on the other hand ensures users' privacy by having each user send his/her data with noise added. Then, not even the central server has access to the ground truth data and must rely on noisy data. For example, in case of LDP on users' height data (Figure 1), the analyzer cannot have precise knowledge of each user's data but may guess the average height accurately with many users.

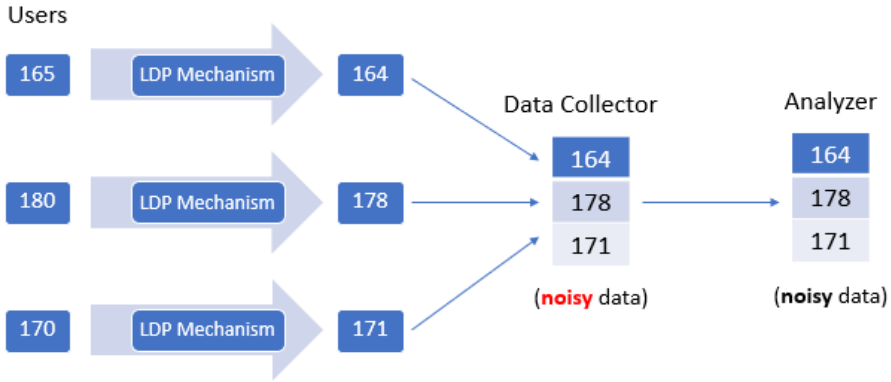


Figure 1: An example of local differential privacy

LDP can be used to analyze decentralized social networks such as epidemic networks. For example, with the advent of COVID19 in 2020, we could have used our limited supply of vaccines to minimize the spread of the disease by analyzing the graph structure of the epidemic network. The next time there is an epidemic outbreak, it would be much better to survey the local citizens on who they usually meet face-to-face, spot the potential "super-spreaders" using metrics that quantify node importance such as coreness, and utilize the limited supply of vaccines based on the analysis. However, citizens may not be fully honest when conducting the survey in fear of leaking sensitive personal information.

In this paper, we propose a new LDP mechanism for decentralized graphs such that node coreness is preserved with strong privacy guarantees.

II. Main Subject

1) Research Planning

Let an adjacency vector denote the vector with the value 1 (value 0) as the i-th element when the i-th corresponding neighbor is connected (is not connected).

For example, the adjacency vector for user 4 in Figure 2 can be expressed as $\boldsymbol{\gamma_4} = (1, 1, 1, 0, 0)$.

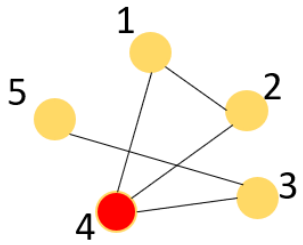


Figure 2: A sample graph

The degree of a node denotes the number of connected neighbors it has. For example, the degree of user 4 in Figure 2 is 3.

A randomized mechanism $\mathcal{M}$ satisfies ϵ-edge LDP if and only if for any two adjacency vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma'}$ that only differ in one bit, and for any $\boldsymbol{s} \in \text{range}(\mathcal{M})$, we have

$$\frac{\Pr\left[\mathcal{M}(\boldsymbol{\gamma}) = s\right]}{\Pr\left[\mathcal{M}(\boldsymbol{\gamma'}) = s\right]} \leq e^{\epsilon}$$

where $\epsilon \geq 0$ implies the privacy budget. Privacy guarantees become tight as the value of ϵ becomes smaller.

The problem statement is as follows. Given n number of users with each user's adjacency vector $\boldsymbol{\gamma_i}$ given as the input, the output must be a synthetic graph G that preserves the coreness (or core numbers) of the original graph as much as possible while satisfying ϵ-edge LDP.

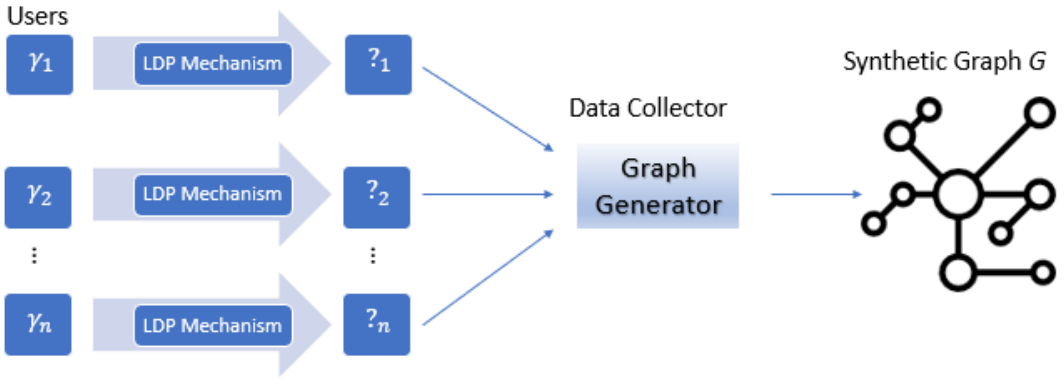


Figure 3: Problem statement

We compared our model to two prior works: Randomized Response (RR) and Degree-Preserving Randomized Response (DPRR).

In the RR model, each bit in each adjacency vector $\boldsymbol{\gamma_i}$ is flipped with probability $p = \frac{1}{e^{\epsilon}+1}$.
A visual depiction is shown below.

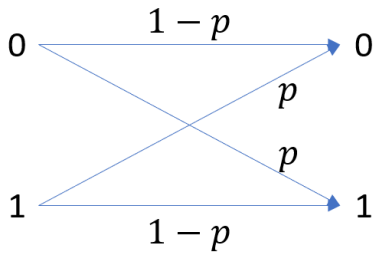


Figure 4: The schematic of Randomized Response (RR)

The problem with RR is that the graph becomes much denser. A common characteristic among all social networks is the sparsity. If each bit of an adjacency vector is flipped with equal probability, the resulting vector will have a much higher degree than it originally had. This leads to too much noise in the process. The high space complexity for graphs with a lot of users is also a problem.

DPRR resolves the issues of RR by using edge sampling after flipping each bit in each adjacency vector $\boldsymbol{\gamma}_i$ with probability $p = \frac{1}{e^{\epsilon}+1}$. Edge sampling is a technique that disconnects some of the edges of a node with a probability derived such that the total degree is maintained. However, the data collector must have prior knowledge of users' degrees. The privacy of users' degrees could be preserved by having the users add Laplace noise to their degree values. Here, one-tenth of the total privacy budget is used on protecting the degree values and the remaining budget is used on protecting the adjacency vectors. A visual depiction is shown below.

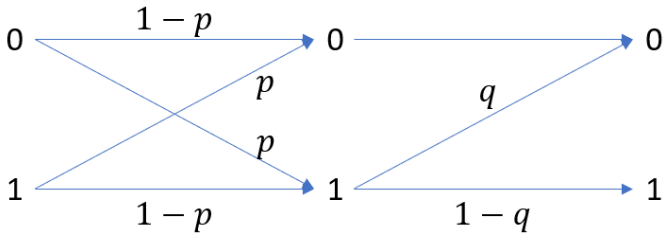


Figure 5: The schematic of Degree-Preserving Randomized Response (RR)

We propose a new mechanism, Degree-Preserving Asymmetric Bit Flipping (DPABF), which preserves user degrees without informing the values to the data collector. In our method, the probability of flipping bit 1 differs from bit 0 such that the degree is preserved.

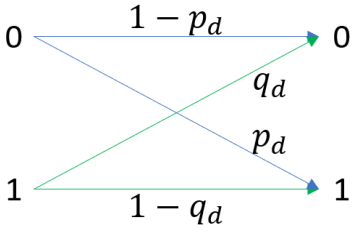


Figure 6: The schematic of our proposed mechanism

Let the adjacency vector and the noisy adjacency vector with degree d be $\overrightarrow{\gamma_d} = (a_1, a_2, \ldots, a_n)$ and $\overrightarrow{\gamma_d}' = (\widetilde{a_1}, \widetilde{a_2}, \ldots, \widetilde{a_n})$, respectively. Let the probability of flipping bit 0 be $p_d = \Pr(\widetilde{a_k} = 1 \mid a_k = 0)$ and the probability of flipping bit 1 be $q_d = \Pr(\widetilde{a_k} = 0 \mid a_k = 1)$, respectively. To preserve the degree of each user after applying edge-LDP, the following equation must hold for $1 \leq d \leq n-1$.

$$d = (n-d)p_d + d(1-q_d)$$

The degree constraint equation reduces to $q_d = \frac{n-d}{d} p_d$.

To satisfy the edge-LDP constraint, the inequalities below must hold for $1 \le d \le n-2$.

$$e^{-\epsilon} \le \frac{1-p_d}{q_{d+1}} \le e^{\epsilon}$$

$$e^{-\epsilon} \le \frac{p_d}{1-q_{d+1}} \le e^{\epsilon}$$

The first approach is to use convex optimization to find the optimal values of p_d and q_d for $1 \le d \le n-1$. Let the objective function that we need to minimize be $L(\boldsymbol{A})$, where $\boldsymbol{A}$ is the adjacency matrix. Then, $L(\boldsymbol{A})$ can be formulated as follows.

$$L(\boldsymbol{A}) = \sum_{d=1}^{\lfloor \frac{n}{2} \rfloor} Var\{\deg(M(\boldsymbol{a}_d)|d)\}$$

Let $B(n,p)$ denote the binomial distribution with n number of trials and probability of success of p.

Since $\deg(M(\boldsymbol{a}_d)|d) \sim \{B(d, 1-q_d) + B(n-d, p_d)\}$,

$$Var\{\deg(M(\boldsymbol{a}_d)|d)\} = q_d(1-q_d) + p_d(1-p_d).$$

Thus, the problem is simplified as

$$minimize\ L(\boldsymbol{A}) = \sum_{d=1}^{\lfloor \frac{n}{2} \rfloor} q_d(1-q_d) + p_d(1-p_d)$$

However, this objective function is unsolvable due to its concavity. The optimal solution to this problem would reside in the vertices of the feasible region, i.e., the corner points of the bound constraints. Since the Hessian matrix of $L(\boldsymbol{A})$ is a negative semidefinite matrix, finding the location of all the corners is NP-hard. Thus, it is difficult to obtain the optimal solution using the corner point method.

To solve this problem in a more tractable form, our next approach was to soften the degree constraint. Let $p_d \approx p_{d+1}$ for $1 \le d \le n-2$. Then the degree constraint becomes

$$q_{d+1} = \frac{n-(d+1)}{d+1} p_{d+1} \approx \frac{n-(d+1)}{d+1} p_d.$$

By substituting q_{d+1} in the edge-LDP constraint, we get

$$p_d = \frac{1}{e^{\epsilon} + \frac{n-(d+1)}{d+1}}, \quad q_d = \frac{n-(d+1)}{n+1} p_d$$

under the assumption that $d \le \left\lfloor \frac{n}{2} \right\rfloor$.

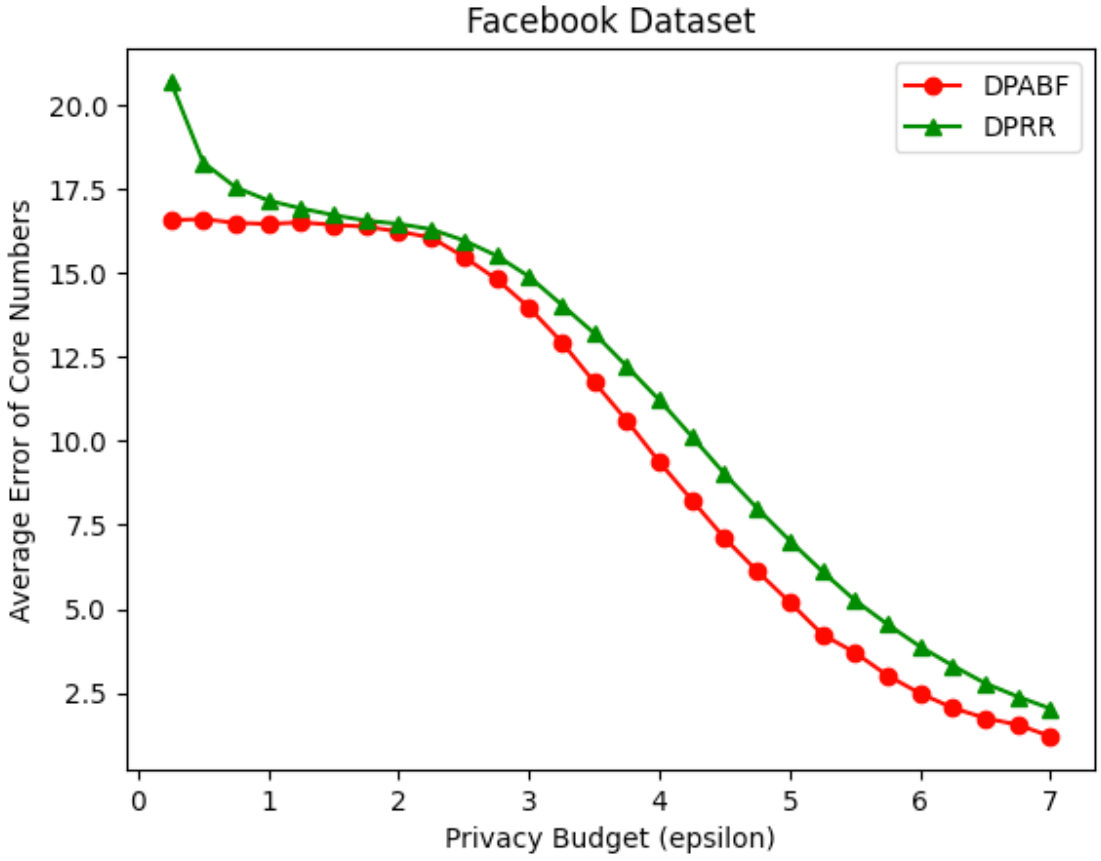
By obtaining a closed form expression of the degree constraint using this method, the probabilities p_d and q_d for $1 \le d \le \left\lfloor \frac{n}{2} \right\rfloor$ become deterministic. A user with degree d may flip his/her adjacency

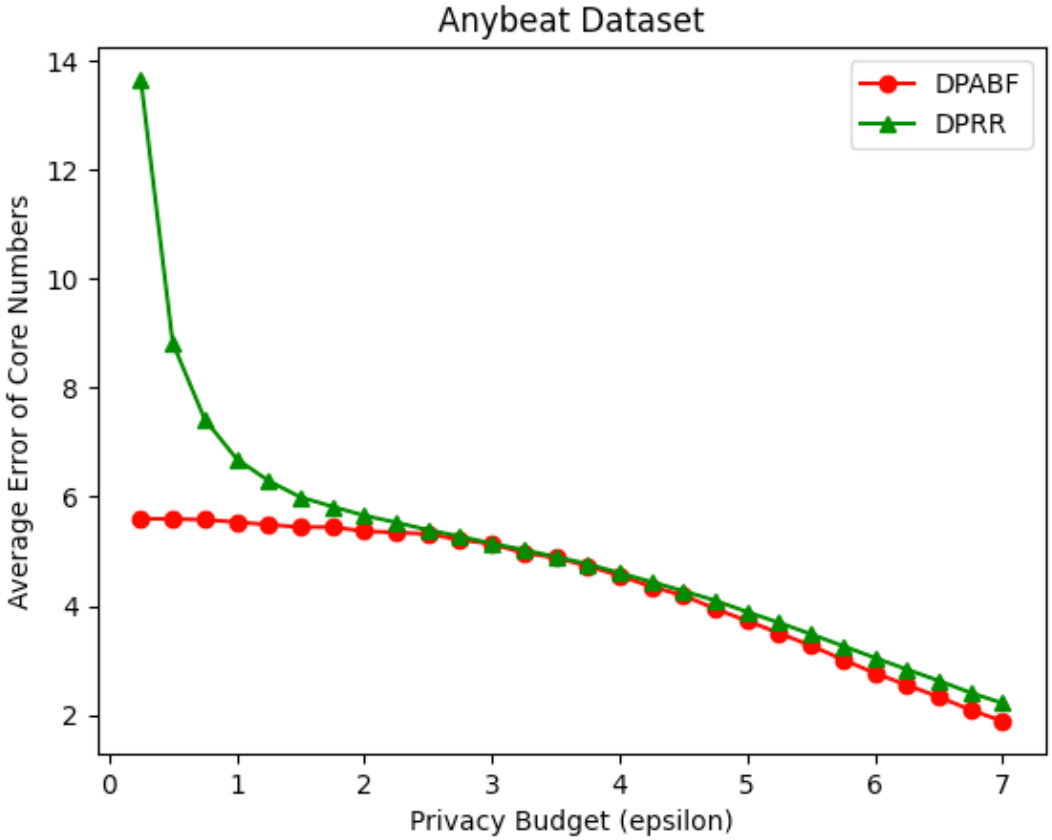
vector by simply calculating p_d and q_d. The data collector cannot know the values of p_d and q_d without prior knowledge of user degrees, only users can. Note that in our approach, the total given privacy budget could be fully used on protecting users' adjacency vectors.

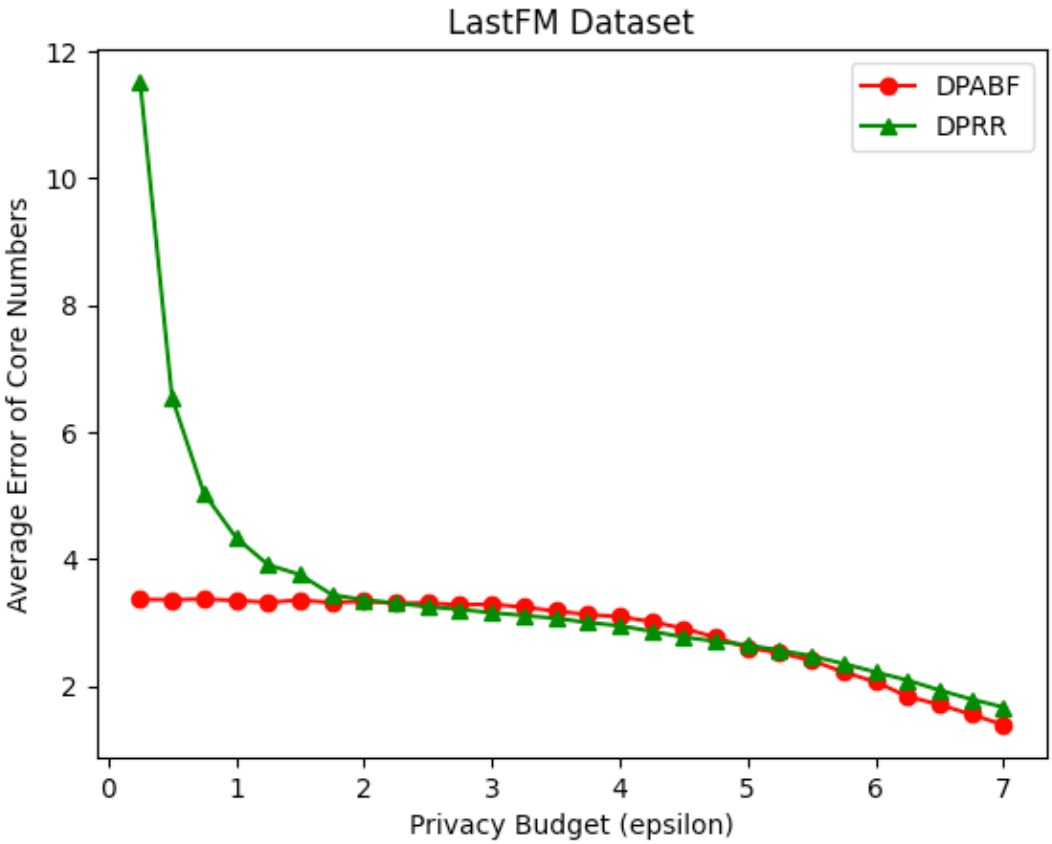
III. Conclusion and Discussion

1) Research Results

Experiments were conducted based on three widely used baseline datasets. The *Facebook* dataset contains 4039 vertices and 88234 undirected edges. The *LastFM Asia Social Network* dataset contains 7624 vertices and 27806 undirected edges. The *Anybeat* dataset contains 12645 vertices and 67.1k undirected edges. For each dataset, the L_1 error of core numbers were computed for values of epsilon from 0.25 to 7 with an increment of 0.25. The coreness (or core number) of a node is a good metric that quantifies the importance of the node in the given graph. Simulation results for Randomized Response were omitted due to its extremely poor performance compared to DPABF and DPRR. The average of the results of 10 separate experiments were plotted as can be seen below.

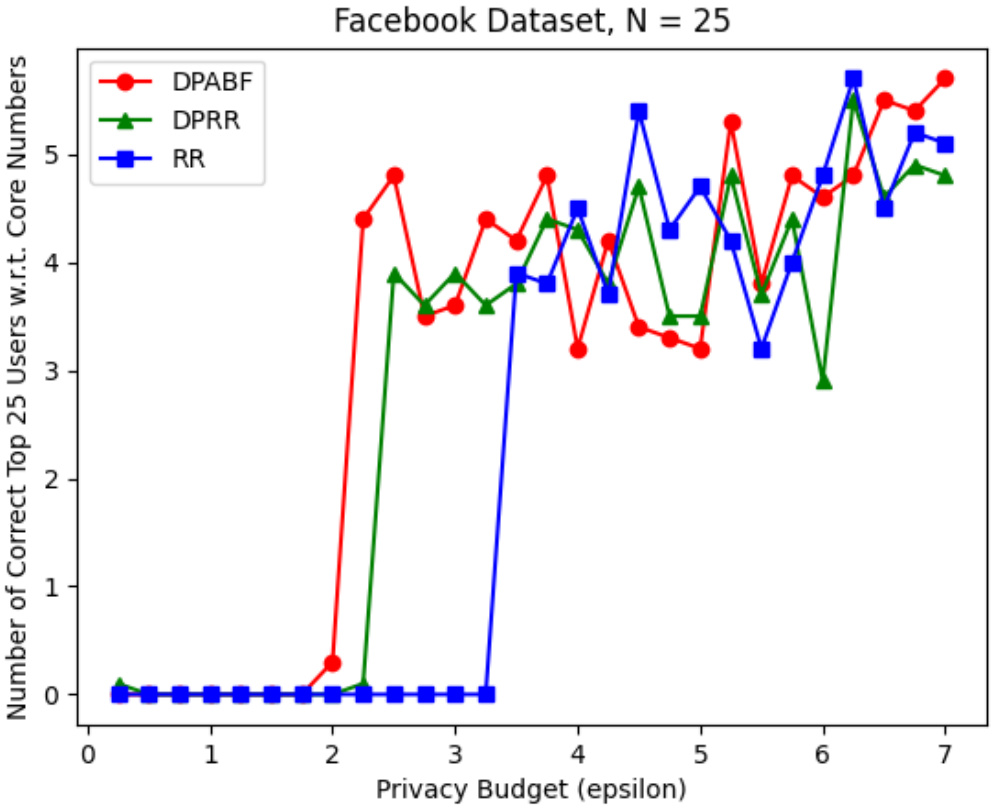


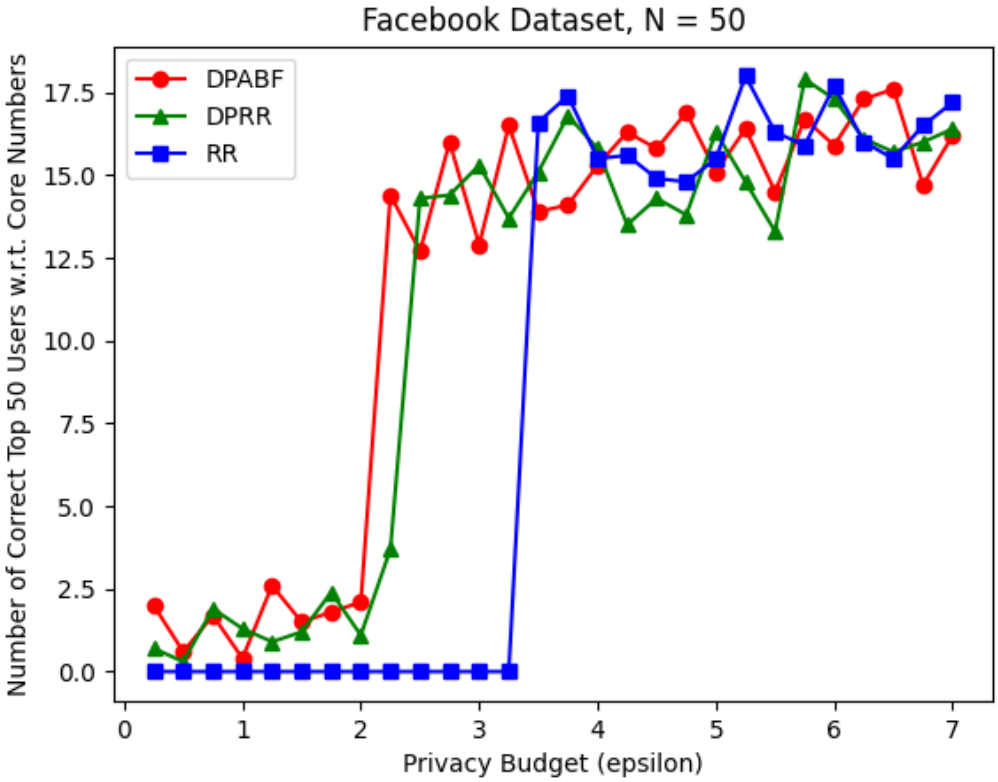


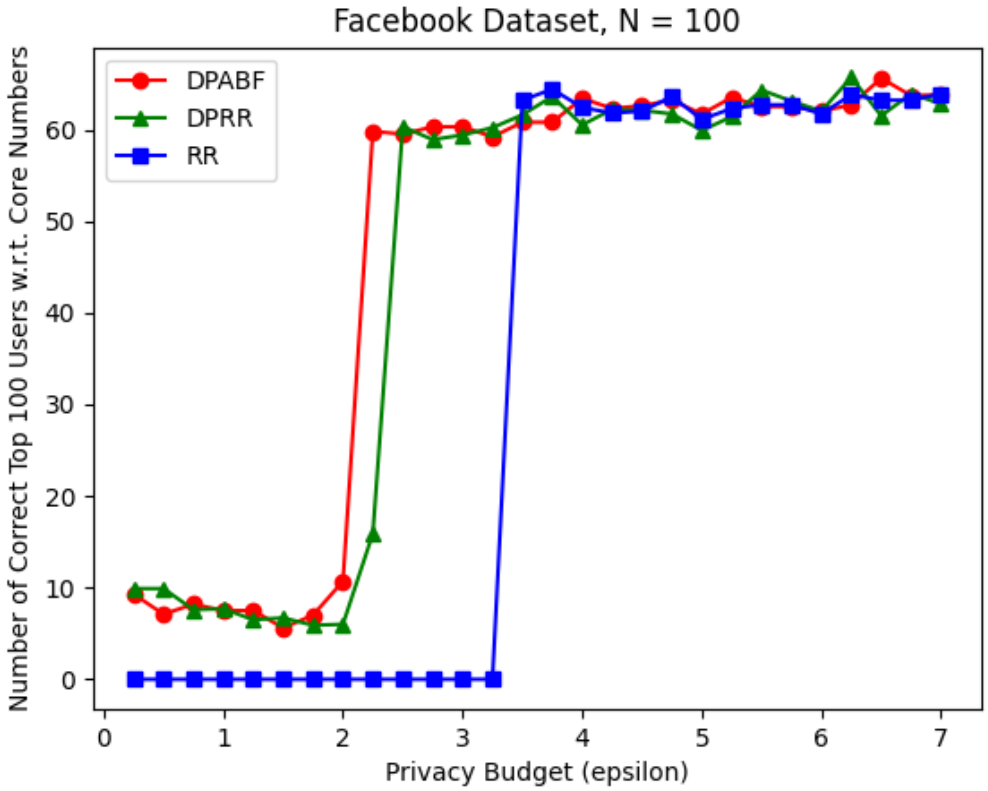


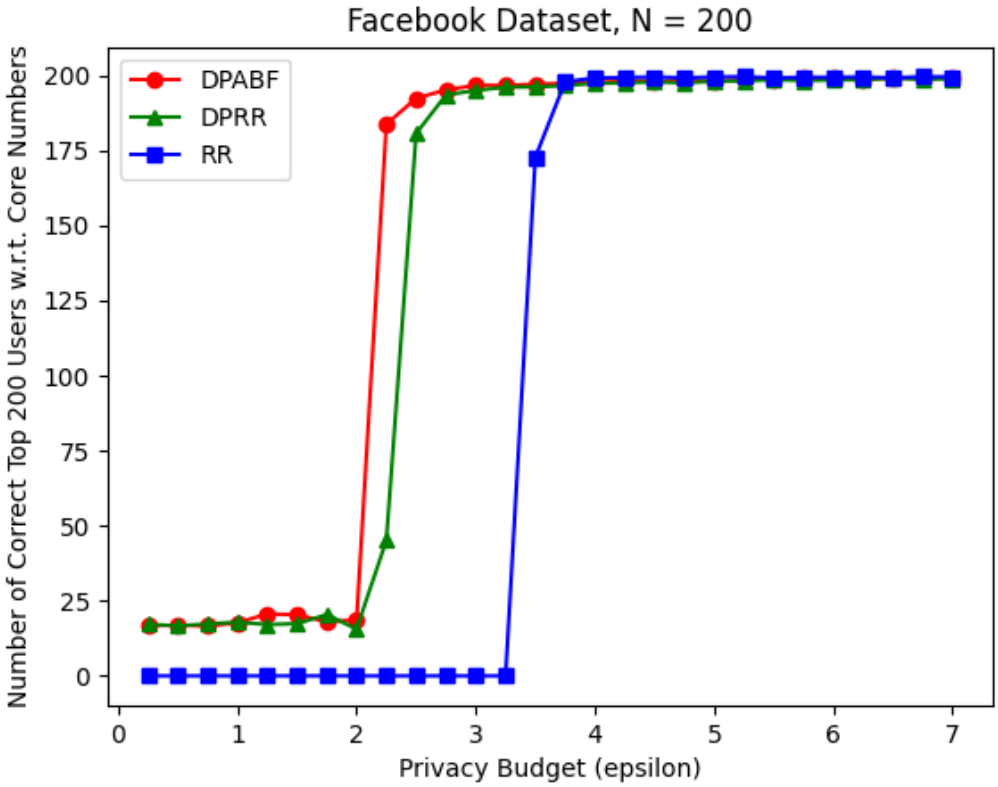
It is empirically shown that in high privacy regions (low privacy budget), our proposed model, DPABF, can retain users' core numbers more accurately compared to DPRR. Experimental results for DPRR diverged as epsilon went to 0, whereas for DPABF it seemed to saturate. In lower privacy regions however, the performance gap seemed too trivial for analysis.

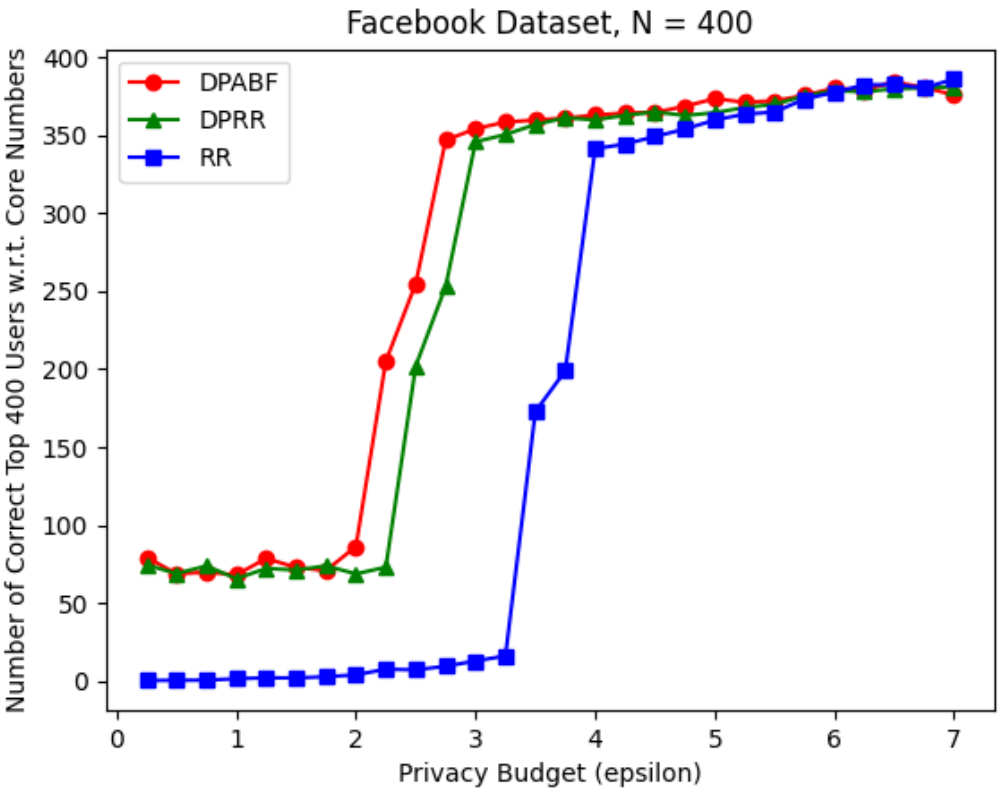
Our next experiment tested how accurately the top N ranked users in terms of coreness values were preserved. In applications such as choosing whom to vaccinate with N limited supply of vaccines, you may only care about who the top N users are, not their exact coreness values. In this scenario, ranking is more important than the absolute coreness values. The *Facebook* dataset was used with values of N ranging from 25 to 400 for this experiment.











2) Discussion

According to the experimental results above, our proposed model can retain users' coreness more accurately in comparison to DPRR in high privacy regions. We conjecture that this is because for low values of epsilon, the Laplace noise added to the node degrees in DPRR significantly increases the error of the degrees, which play a key role in the overall graph structure. Our proposed mechanism, on the other hand, will preserve users' degrees in a probabilistic sense for any given privacy budget, epsilon. We believe that the performance gap between our method and DPRR is distinct due to the fact that our method doesn't split up and allocate the privacy budget to the adjacency vectors and degree values, but instead fully utilizes the whole budget to the adjacency vectors.

Note that when we softened the degree constraint, we chose $p_d \approx p_{d+1}$ over $q_d \approx q_{d+1}$. This is because adjacent vectors of social networks are generally sparse. With more elements of 0 than 1 in any given adjacency vector, p_d doesn't vary with d as severely as q_d does. Therefore, the inevitable error that arises from softening the degree constraint could be lowered by choosing $p_d \approx p_{d+1}$ instead of $q_d \approx q_{d+1}$.

Overall, our proposed model, DPABF is much better than DPRR in low privacy budgets. In higher privacy regions, not so much. We believe this is because the Laplace noise added to the node degrees in DPRR isn't relatively noticeable with high values of epsilon.

An interesting phenomenon that occurred in the *Top N* experiments is the abrupt performance improvement for a specific value of epsilon. The value of epsilon for which the accuracy spiked differed for DPABF, DPRR, and RR. The number N also affected the location of the cliffs on the plot. Nonetheless, our proposed method, DPABF, surpassed DPRR and RR in this experiment. For low values of N, it looks like 10 separate experiments weren't enough to have a smooth plot, but as N increases, you can see that the top N ranked users were quite robust to LDP mechanisms. The analysis of the cliff phenomenon in random graphs is left for future work.

※ References

1. S. L. Warner 1965, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", Journal of the American Statistical Association, vol. 60, p.63-69.

2. S. Hidano, T. Murakami 2023, "Degree-Preserving Randomized Response for Graph Neural Networks under Local Differential Privacy". https://arxiv.org/abs/2202.10209

3. P. M. Pardalos, S. A. Vavasis 1991, "Quadratic programming with one negative eigenvalue is NP-hard", Journal of Global Optimization **1**: 15-22.