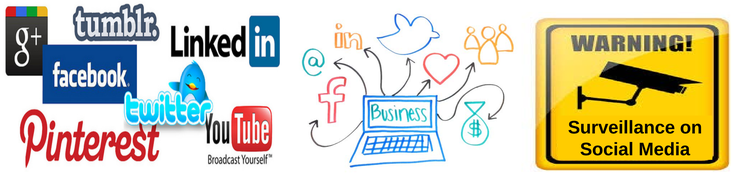
Correlated Network Data Publication via Differential Privacy

Rui Chen[¶], Benjamin C. M. Fung[*], Philip S. Yu[+], Bipin C. Desai [§]

[¶]Hong Kong Baptist University, Hong Kong
[*]McGill University, Montreal, Canada
[+]University of Illinois at Chicago, IL, USA
[§] Concordia University, Montreal, Canada

Introduction

In the last few years, information networks in various application domains, such as social networks, communication networks and transportation networks, have experienced vigorous developments, and have enabled a wide spectrum of data analysis tasks.



Research on social networks | **Targeted advertisements** | **Government surveillance**

Figure 1. Examples of increasing demands on network data

Network data often contains sensitive information. Thus, publishing network data with provable privacy guarantees is of utmost importance. In this paper, we study the problem of releasing possibly correlated network data under *differential privacy* (DP) [1]. We analyze the privacy guarantee of DP in the correlated setting. We propose the first practical *non-interactive* solution for network data publication. Extensive experimental results demonstrate that our solution preserves high utility and scales to large-scale real-life data.

DP under Correlation

Neighboring graphs are defined as graphs differing in at most an edge. *Edge differential privacy* [1, 2] is then defined as follows.

Definition 1. A randomized algorithm $\mathcal{A}$ gives ε-differential privacy if for *any* neighboring graphs G and G' and, for *any* possible output $O \in Range(\mathcal{A})$,

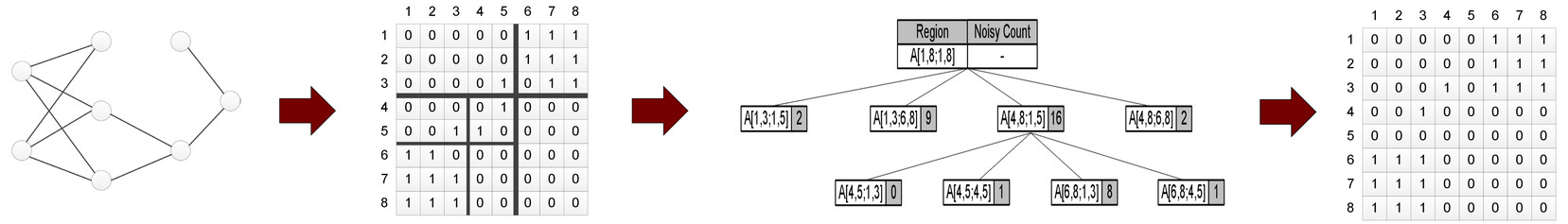
$$Pr[\mathcal{A}(G) = O] \leq \exp(\varepsilon) \times Pr[\mathcal{A}(G') = O]$$

Intuitively, edge DP prevents any single edge from being unveiled. DP gives one of the strongest privacy guarantees; however, its application to network data is hindered by the fact that network data may be inherently correlated [3]: the presence or absence of an edge can be reflected by several other edges. Fortunately, DP is flexible to cope with correlation as long as the extent of correlation can be measured.

Key Observation. To cancel out the effect of data correlation, one should add extra Laplace noise that is proportional to the extent of correlation.

This observation allows DP under correlation to be interpreted in terms of *group differential privacy* [1].

Anonymization Algorithm



Raw graph *G* with a random vertex labeling

Step 1: Identify a good vertex labeling that makes the corresponding adjacency matrix form dense clusters of 1s by a greedy permutation-based algorithm

Step 2: Explore dense/sparse regions of the adjacency matrix by adapting a standard quadtree structure in a differentially private and data-dependent way

Step 3: Reconstruct a noisy adjacency matrix based on a novel use of the exponential mechanism

Experimental Evaluation

We employ four real-life datasets that are publicly available and examine data utility of sanitized network data in terms of three common data analysis tasks: *cut query*, *degree distribution* and *shortest path length*.

We compare our density-based exploration and reconstruction approach (*DER*) with a random graph (*Random*) of the same numbers of nodes and edges, a simple Laplace mechanism based approach (*Laplace*) [4] and a simplified version of *DER* (*DE*, without Step 3).

Table 1. Experimental dataset statistics

Datasets	|V|	|E|	Edge Density
ca-GrQc	5,242	14,484	0.00106
ca-HepTh	5,000	17,120	0.00137
wiki-Vote	7,115	100,762	0.00398
STM	1,012	7,860	0.01536

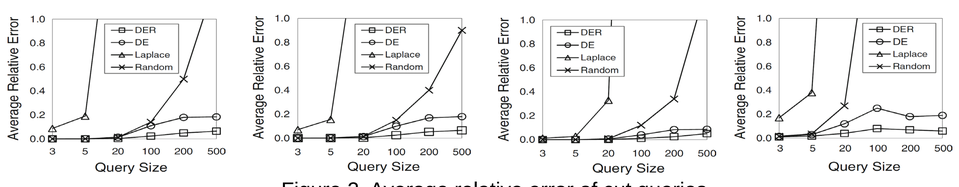


Figure 2. Average relative error of cut queries

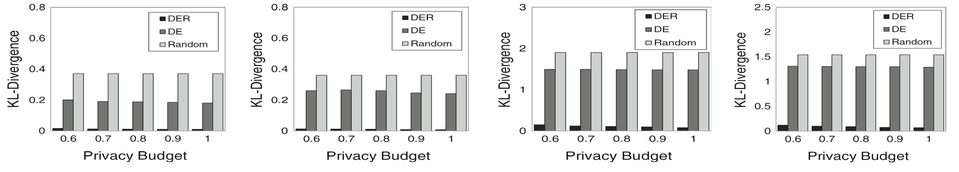


Figure 3. KL-divergence of degree distributions

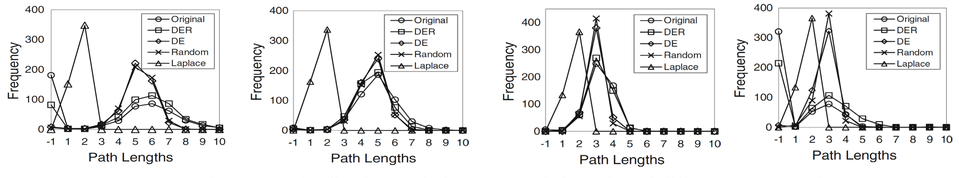


Figure 4. Distributions of shortest path lengths of different approaches

Conclusions

We analyzed the properties of DP in the correlated setting and indicated that DP is flexible to handle data correlation. We presented an efficient non-interactive approach for publishing correlated network data. This is the first work that gives a practical solution for network data publication under DP. Extensive experiments demonstrate that our solution performs well for various data analysis tasks on different types of real-life network datasets.

References

1. C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of TCC*, pp. 265-284, 2006.
2. M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In Proc. of ICDM, pp. 169-178, 2009.
3. D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Proc. of SIGMOD*, pp. 193–204, 2011.
4. A. Gupta, A. Roth, and J. Ullman. Iterative constructions and private data release. In *Proc. of TCC*, pp. 339-356, 2012.