Privacy-Preserving Techniques for Social Network Analysis

André L. C. Mendonça, Felipe T. Brito, Javam C. Machado

Laboratório de Sistemas e Banco de Dados (LSBD)
DC/UFC – CEP 60440-900 – Fortaleza – CE – Brazil

{andre.luis,felipe.timbo,javam.machado}@lsbd.ufc.br

Abstract. *With the increasing concerns over data privacy, preserving the privacy of individuals in social network analysis has become crucial. This tutorial provides a comprehensive overview of methods and techniques to protect individual privacy while conducting social network analysis. We perform a deep analysis of differential privacy, which is a rigorous mathematical framework to protect individual privacy while enabling accurate analysis of network structure and characteristics. Additionally, this tutorial explores a variety of examples and case studies to demonstrate the application of these techniques in practical scenarios.*

Introduction

Social network analysis [Knoke and Yang 2019] is a powerful tool that allows the investigation of human interactions, the discovery of hidden patterns, and the gain of valuable insights into social behaviors. It enables the understanding of how ideas are spread, communities are formed, and relationships shape individuals' lives. However, as we get deeper into the field of social network analysis, we can not overlook the critical privacy issues and concerns that arise. With the vast amount of personal data shared within social networks, it is crucial to recognize the importance of privacy in this context. Privacy serves as a shield that ensures the individual's autonomy, protects their sensitive information, and upholds ethical principles [Kearns and Roth 2019]. By prioritizing privacy in social network analysis, we can create a secure and trustworthy environment that respects users' rights, preserves their confidentiality, and ensures responsible data practices by analysts.

In the last decades, many privacy techniques were designed with their own requirements to protect individuals' privacy, such as *k-anonymity*, *l-diversity*, *t-closeness*, *δ-presence*, among others [Brito and Machado 2017]. However, all of these approaches assume that a malicious user has limited background knowledge, which is not true in real-world scenarios. In this context, differential privacy (DP) [Dwork 2006] emerges as a crucial concept. It acts as a balance between extracting meaningful insights from data and preserving individuals' privacy. Differential privacy has been applied in the industry by companies such as Apple, Google, Uber and also in the public sector by U.S. agencies, such as the U.S. Census Bureau. DP also provides a rigorous mathematical framework that allows the analysis of aggregate data while protecting the sensitive information of individuals. Consequently, it is also extensively applied in social network analysis.

Many efforts have been made towards differentially private approaches for social network analysis over the past years [Silva et al. 2017, Xia et al. 2021, Farias et al. 2023, Brito et al. 2023]. In this context, two main types of DP are particularly relevant: edge

differential privacy (edge-DP) [Hay et al. 2009] and node differential privacy (node-DP) [Kasiviswanathan et al. 2013]. The essential difference between them lies in the definition of neighboring graphs. In the standard DP model, two databases are neighbors if they differ by at most one record. In the graph context, edge-DP describes two graphs as neighbors if they differ on a single edge. On the other hand, node-DP defines a pair of graphs to be neighbors if they differ by exactly one node and its incident edges. Intuitively, edge differential privacy ensures that the output of a DP algorithm does not reveal the inclusion or removal of a particular edge in the graph, while node differential privacy hides the inclusion or removal of a node together with all its adjacent edges. Both edge and node differential privacy play important roles in preserving privacy while allowing meaningful analysis of social network data.

In this context, there are two main approaches to protecting social network information from an analyst (or a malicious user) who has access to sensitive information. The first one is by answering the analyst's queries in a private manner so that the existence of any social network information is almost indistinguishable before and after seeing the private query outputs. The second approach is releasing an entire social network that is a close approximation to the true one but is guaranteed to be private according to differential privacy, so the analyst's queries will be answered using the released network. In this tutorial, we explore the application of various techniques existing in the literature that apply both approaches.

Main topics

The tutorial intends to discuss the following topics:

Social Network Analysis Basics: This section presents an overview of the social network data structures. In general, social networks are mainly composed of a set of nodes (also known as vertices) and edges, where nodes concern entities of interest, while edges represent relationships among entities. Besides, we discuss the social network's different representations. Commonly, social networks are represented by graphs and adjacency matrices. Finally, this section also presents common multipurpose algorithms and metrics for social network analysis, such as degree distribution, subgraph counting, clustering coefficients, centrality measures, among others.

Privacy Threats in Social Networks: This section discusses the importance of privacy in social network analysis by presenting privacy breaches and concerns in social network analysis. We argue that a privacy attack involves combining auxiliary information with de-identified data to re-identify individuals. We study the differences between the main privacy attacks: identity disclosure, attribute disclosure, link disclosure, and graph metrics disclosure. The privacy issues are also motivated by presenting real scenarios in which sensitive personal information is identified in social network analysis. This section finishes by discussing the trade-off between privacy and utility in social network data sharing.

Differential Privacy Fundamentals: Differential privacy is a formal privacy model originally designed for use on raw data in order to provide robust privacy guarantees without depending on an adversary's background knowledge. The main idea behind DP is that a given query is answered by a *randomized algorithm* (also referred to as a mechanism) that queries the private information and returns a randomized answer

sampled from an *output distribution*. Because it offers a mathematically rigorous method of ensuring privacy, it has become the de facto standard for private data release. In this section, we describe the main concepts of differential privacy, stating the definition of neighboring datasets along with their applicability. Additionally, we present the privacy budget setting (parameter ϵ), which is responsible for balancing the trade-off between privacy and the utility of the results. Finally, we introduce the Laplace and the exponential mechanisms, which are methods to achieve differential privacy, exhibiting interesting DP properties.

Privacy-Preserving Social Network Analysis: This section presents differentially private anonymization techniques for social network data. These anonymization techniques aim to preserve individuals' privacy while sharing social network data for multipurpose analysis (e.g., community detection, link prediction, subgraph counting, and centrality measures). Additionally, this section presents the definition of differential privacy for social network data. Overall, the application of differential privacy on network data and raw data differs in terms of privacy preservation goals, data representation, perturbation techniques, and the types of analysis or queries that are protected. We formalize the notions of edge-DP and node-DP, variants of differential privacy specifically designed for protecting the privacy of network data. We also present methods that post-process the data to boost the accuracy of existing differentially private algorithms. The section finishes by presenting mechanisms applicable to the different notions of differential privacy in social networks.

Practical Implementation of Differential Privacy on Social Networks: This section presents a variety of differentially private mechanisms implementation for different purpose tasks in social networks. We perform a practical demonstration using node and edge differential privacy-based approaches for a variety of social network analysis tasks, like degree distribution, subgraph counting, diameter, centrality measures, among others. We adopt the DPGraph [Xia et al. 2021], a benchmark platform for differentially private graph analysis, which enables users to understand the trade-off between privacy, accuracy, and performance of existing work and discover suitable algorithms for their applications. Case studies and real-world examples are used to consolidate the presented approaches for the different tasks.

Future Directions and Open Challenges: This section presents the current research trends in differential privacy on social networks. It also discusses some newly advanced topics, such as dynamic and evolving social networks, and social network-based machine learning. In this section, we also present our current studies in the field of differentially private mechanisms for weighted networks, networks with opt-in and opt-out users, and attributed social networks. This section finishes by presenting existing limitations and open challenges while motivating with potential future directions in the field of privacy-preserving social network analysis. We argue that getting DP to work in practice requires a team of experts and that the community needs more examples of real-world deployments.

About the authors:

André L. C. Mendonça is a Computer Science Ph.D. student at the Universidade Federal do Ceará (UFC), with a 3-months sandwich period at Laboratoire d'Informatique de

Grenoble (LIGLab), Grenoble, France. He obtained an M.Sc. and B.Sc. in Computer Science from UFC. He is a researcher at the Laboratório de Sistemas e Banco de Dados (LSBD/UFC), currently working with differential privacy mechanisms for attributed graphs. He is interested in the following research topics: Data Privacy, Differential Privacy, Graphs, and Social Networks. He has over 6 years of experience in the field of privacy, with papers published in this area at important conferences.

Felipe T. Brito has a Ph.D. in Computer Science from Universidade Federal do Ceará (UFC), with a sandwich year at AT&T Labs, New York, USA. He obtained an M.Sc. and a B.Sc. in Computer Science, also from UFC. He is a researcher at the Laboratório de Sistemas e Banco de Dados (LSBD/UFC), currently working with differentially private mechanisms for weighted graph data structures. Felipe has over 8 years of experience in the field of privacy, with important papers published in this area, including both VLDB and SIGMOD conferences. His topics of interest also include machine learning and data management.

Javam C. Machado is a full professor at the Computer Science Department of Universidade Federal do Ceará (UFC), Brazil. He obtained an M.Sc. in Computer Science from the Universidade Federal do Rio Grande do Sul, Brazil, and a Ph.D. degree in Computer Science from the Université de Grenoble, France. In 2010, Javam started the Laboratório de Sistemas e Banco de Dados (LSBD) and has coordinated it since then. For 12 years, Javam was the UFC director of information technology; for 2 years, he served as the coordinator of research and technological innovation also at UFC. He was the coordinator of the SBC Special Database Commission (2017) and visiting researcher at TelecomSud-Paris – FR (2001) and at AT&T Labs-Research – USA (2018; 2020; 2023). Javam has published more than 170 scientific papers and has advised 40 M.Sc. and 5 Ph.D. students. A CNPQ researcher, Professor Javam is scientifically interested in the areas of data privacy and non-discrimination in machine learning techniques.

References

Brito, F. T., Farias, V. A., Flynn, C., Majumdar, S., Machado, J. C., and Srivastava, D. (2023). Global and local differentially private release of count-weighted graphs. *Proceedings of the ACM on Management of Data*, 1(2):1–25.

Brito, F. T. and Machado, J. C. (2017). Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. *Jornadas de atualização em informática*, pages 91–130.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Farias, V. A., Brito, F. T., Flynn, C., Machado, J. C., Majumdar, S., and Srivastava, D. (2023). Local dampening: Differential privacy for non-numeric queries via local sensitivity. *The VLDB Journal*, pages 1–24.

Hay, M., Li, C., Miklau, G., and Jensen, D. (2009). Accurate estimation of the degree distribution of private networks. In *International Conference on Data Mining*, pages 169–178. IEEE.

Kasiviswanathan, S. P., Nissim, K., Raskhodnikova, S., and Smith, A. (2013). Analyzing graphs with node differential privacy. In *Theory of Cryptography Conference*, pages 457–476. Springer.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Knoke, D. and Yang, S. (2019). *Social network analysis*. SAGE publications.

Silva, R. R. C., Leal, B. C., Brito, F. T., Vidal, V. M., and Machado, J. C. (2017). A differentially private approach for querying rdf data of social networks. In *International Database Engineering & Applications Symposium*, pages 74–81.

Xia, S., Chang, B., Knopf, K., He, Y., Tao, Y., and He, X. (2021). Dpgraph: A benchmark platform for differentially private graph analysis. In *International Conference on Management of Data*, pages 2808–2812.