Differentially Private Release of Count-Weighted Graphs

Felipe T. Brito, Javam C. Machado

Universidade Federal do Ceará (UFC)
Fortaleza, CE, Brasil

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

Abstract. *This work proposes many contributions to privacy in complex systems, mainly ones modeled as count-weighted graphs. As graph data usually contain users' sensitive information, preserving privacy when releasing this type of data becomes a crucial issue. In this context, differential privacy (DP) has become the de facto standard for data release under strong mathematical guarantees. However, various challenges persist in effectively implementing DP to graph data, including balancing privacy protection with data utility and scalability concerns. To bridge these gaps, we propose several efficient techniques and approaches to release graph data while maintaining a robust level of privacy protection. Our results were published in the top-tier venues in the field of data management. Additionally, we disseminated our knowledge and expertise obtained during this Ph.D. research through tutorials and short courses presented at both national and international conferences.*

Resumo. *Este trabalho propõe várias contribuições para a privacidade em sistemas complexos, principalmente aqueles modelados como grafos ponderados por contagem. Como dados de grafos geralmente contêm informações sensíveis dos usuários, preservar a privacidade ao compartilhar esse tipo de dado se torna uma questão crucial. Nesse contexto, a privacidade diferencial (PD) tornou-se o padrão para o compartilhamento de dados com fortes garantias matemáticas. No entanto, diversos desafios persistem na implementação eficaz de PD em dados de grafos, incluindo o equilíbrio entre proteção de privacidade, utilidade dos dados e preocupações com escalabilidade. Para preencher essas lacunas, propomos várias técnicas e abordagens eficientes para liberar dados de grafos, mantendo um nível robusto de proteção de privacidade. Nossos resultados foram publicados nos principais veículos da área de gerenciamento de dados. Além disso, disseminamos o conhecimento e a expertise obtidos durante esta pesquisa de doutorado por meio de tutoriais e minicursos apresentados em conferências nacionais e internacionais.*

1. Thesis Defense Data

- Defense date: August 3rd, 2023.
- Graduate Program: Programa de Pós Graduação em Ciência da Computação (MDCC) - Universidade Federal do Ceará (UFC).
- Category: Doctorate degree.
- Committee: Cesar Lincoln Cavalcante Mattos (Universidade Federal do Ceará), Carlos Eduardo Santos Pires (Universidade Federal de Campina Grande), Mirella Moura Moro (Universidade Federal de Minas Gerais), and Divesh Srivastava (AT&T Chief Data Office - USA).

2. Context and Problem

Graphs are a fundamental tool for understanding the structure and behavior of complex systems [Newman 2003]. In this context, *count-weighted graphs* refer to a graph type in which each edge is assigned a weight indicating the frequency of its occurrence within a dataset. Over the past decade, they have been widely adopted to characterize complex systems in the real world, such as targeted marketing and advertising [Leskovec et al. 2007], information campaigns through social media [Ferrara et al. 2016, Varol et al. 2017], analysis of influential people and the interactions between them [Camacho et al. 2020], propagation-based fake news detection [Matsumoto et al. 2021], spread of epidemic disease [Manríquez et al. 2021], among others [Brito 2023]. Because count-weighted networks contain sensitive information [Brito et al. 2023], releasing this type of data for data scientists, researchers, machine learning engineers, and other stakeholders who deal with sensitive data, may seriously jeopardize individuals' privacy. Current laws and regulations on data privacy require that individuals are no longer re-identifiable from released information [Brazil 2018, European Union 2016, Federal Communications Commission 2018]. In this context, a new paradigm has emerged: differential privacy (DP) [Dwork 2006].

Differential privacy is a formal privacy model originally designed for use on raw data to provide robust privacy guarantees without depending on an adversary's background knowledge. DP is not a single tool, but rather a paradigm that quantifies and manages privacy violation risks. DP can be applied from simple statistical estimations to graph analytics and machine learning [Brito et al. 2024]. Over time, differential privacy has come to be considered in graph analytics [Brito et al. 2024]. 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]. However, when graphs have weighted edges, neither edge-DP nor node-DP models offer appropriate privacy guarantees [Brito 2023]. Then, we investigate the problem of releasing count-weighted graphs under differential privacy guarantees.

3. Goal

In our thesis, we focus on proposing techniques to achieve DP in complex systems, mainly ones modeled as count-weighted graphs. At the same time, we aim to preserve the original graph's characteristics while ensuring the released data retains its utility and accuracy, allowing data scientists, researchers, machine learning engineers, and other stakeholders to perform meaningful analyses without compromising privacy. This is particularly challenging because ensuring data privacy often involves adding noise to the data, which can distort the original structure, weights, and relationships in the graph. Balancing privacy with data utility requires sophisticated techniques to minimize information loss while still providing robust privacy guarantees. Additionally, achieving this balance becomes even more complex in terms of scalability, as the methods must efficiently handle large-scale graphs without significantly increasing computational overhead or processing time.

4. Contributions

The major contributions of this thesis emerge from a comprehensive examination of state-of-art, the proposal of scalable differentially private techniques, meaningful analysis from the achieved results, and the dissemination of knowledge in the data privacy field.

Surveys on data privacy in the Portuguese language: When we started this Ph.D. research, literature on data privacy, particularly in Portuguese, was scarce. To address this, we proposed two surveys: one introducing the main concepts of data privacy preservation [Brito and Machado 2017] and another describing techniques to prevent reidentification in graph data [Mendonça et al. 2024].

New neighboring weight graphs definitions with unknown graph topology: Existing DP techniques assume the graph topology is known, resulting in perturbation only of the edge weights, which can cause excessive weight value distortion. To address this limitation, we propose new definitions for neighboring weight graphs with unknown topology [Brito et al. 2023].

Scalable and accurate DP techniques for releasing count-weighted graphs: Having established DP to address count-weighted graphs with unknown topology, we could apply this new notion to perturb an input graph and produce a noisy version that is capable of handling large-sized graphs, since we achieved overall complexity of $O(|E|.log(|E|))$ [Brito et al. 2023].

Identifying influential nodes with differential privacy: We also proposed a differentially private algorithm that chooses iteratively nodes that maximize influence metrics [Farias et al. 2020, Farias et al. 2023].

New private count-weighted network datasets available for download: In the literature, there is a notable scarcity of datasets representing count-weighted graphs. Taking this into account, we pre-processed eight real-world network datasets from different domains and characteristics and released them at kaggle[1], the world's largest data science community.

Tutorials covering private social network analysis and graph analytics: Aiming to share our knowledge and expertise obtained during this Ph.D. research, we also proposed two tutorials [Brito et al. 2024, Mendonça et al. 2023] that reviewed over 100 references and explored a set of DP methods and techniques applicable to graph analytics and social networks.

Exploring differential privacy to address additional problem domains: Finally, throughout this Ph.D. research, we acquired valuable insights that have enabled us to contribute to further publications within the differential privacy field [Leal et al. 2018, Mendonça et al. 2017, Monteiro et al. 2023, Neto et al. 2018, Silva et al. 2017].

5. Advancement in the state of the art

Sealfon [Sealfon 2016] introduced DP in the context of neighboring weight functions and proposed the notion of edge weight-DP. This assumption was recently adopted by several authors in the literature [Chen et al. 2022, Fan and Li 2022, Li et al. 2017, Pinot et al. 2018, Wang and Long 2019]. Under this model, the graph topology is assumed to be public and the private information to protect are only the edge weights. However, it may be desirable in many applications to handle count-weighted graph release when the topology is unknown, for example, when protecting the presence or absence of interactions in a human contact network, or the existence or absence of phone calls, text messages, or emails exchanged between people [Brito 2023]. Unlike previous studies, we

[1] https://www.kaggle.com/datasets/felipetimbo/weighted-networks/

were the first to introduce a new concept of edge-weight differential privacy and propose several methods to ensure the privacy of both the graph structure and edge weights.

6. Evaluation

We compared our approaches against six competing methods: geometric mechanism [Ghosh et al. 2012], exponential mechanism [McSherry and Talwar 2007], log-laplace [Le Ny and Pappas 2013], truncation [Hardt and Roth 2012], high-pass filter [Cormode et al. 2012] and priority sampling [Duffield et al. 2007]. We also utilized the following statistics to measure the utility of the released graph: graph similarity, KL divergence between node degree and edge weight distributions, absolute error in the sum of edge weights, average weighted shortest paths, clustering coefficient, relative error in node strength, influence, and weighted PageRank, among others. We showed that the low-time complexity obtained by our work enabled our approaches to handle large-size graphs and provided high-utility graph analytics.

7. Scientific Production

Felipe T. Brito, Victor A. E. Farias, Cheryl Flynn, Subhabrata Majumdar, Javam C. Machado, Divesh Srivastava. Global and Local Differentially Private Release of Count-Weighted Graphs. *Proceedings of the ACM on Management of Data* [Brito et al. 2023].

Victor A. E. Farias, **Felipe T. Brito**, Cheryl Flynn, Javam C. Machado, Subhabrata Majumdar, Divesh Srivastava. Local Dampening: Differential Privacy for Non-numeric Queries via Local Sensitivity. *Proceedings of the VLDB Endowment* [Farias et al. 2020].

Felipe T. Brito, André L. C. Mendonça, Javam C. Machado. A Differentially Private Guide for Graph Analytics. *Proceedings of the 27th International Conference on Extending Database Technology* [Brito et al. 2024].

Victor A. E. Farias, **Felipe T. Brito**, Cheryl Flynn, Javam C. Machado, Subhabrata Majumdar, Divesh Srivastava. Local Dampening: Differential Privacy for Non-numeric Queries via Local Sensitivity. *The VLDB Journal* [Farias et al. 2023].

Felipe T. Brito, Javam C. Machado. Preservação de Privacidade de Dados: Fundamentos, Técnicas e Aplicações. *Jornadas de atualização em informática* [Brito and Machado 2017].

André L. C. Mendonça, **Felipe T. Brito**, Javam C. Machado. Privacy-Preserving Techniques for Social Network Analysis. *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados* [Mendonça et al. 2023].

André L. C. Mendonça **Felipe T. Brito**, Javam C. Machado. Análise de Dados Privada em Redes Sociais. *Jornadas de atualização em informática* [Mendonça et al. 2024].

Rôney Reis C. Silva, Bruno C. Leal, **Felipe T. Brito**, Vânia M. P. Vidal, Javam C. Machado. A Differentially Private Approach for Querying RDF Data of Social Networks. *Proceedings of the 21st International Database Engineering & Applications Symposium.* [Silva et al. 2017].

Felipe C. Monteiro, **Felipe T. Brito**, Iago C. Chaves, Javam C. Machado. Compartilhamento de Dados de Tráfego de Rede Utilizando Privacidade Diferencial. *Anais do L Seminário Integrado de Software e Hardware* [Monteiro et al. 2023].

Eduardo R. D. Neto, André L. C. Mendonça, **Felipe T. Brito**, Javam C. Machado. PrivLBS: Uma Abordagem para Preservação de Privacidade de Dados em Serviços Baseados em Localização. *Anais do XXXIII Simpósio Brasileiro de Banco de Dados.* [Neto et al. 2018].

Acknowledgments

This work was partially done while the author was a visiting research scientist at AT&T Labs, USA. This research was supported by CAPES/Brazil under grant numbers 189723/2018-01, and 454577/2018-01, and also by CNPq/Brazil under grant number 316729/2021-3. Additional funding was provided by Laboratório de Sistemas e Banco de Dados LSBD/UFC.

References

Brazil (2018). Lei Geral de Proteção de Dados Pessoais. `http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm`. Online; accessed 15 May 2023.

Brito, F. T. (2023). *Differentially private release of count-weighted graphs*. PhD thesis, Universidade Federal do Ceara.

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.

Brito, F. T., Mendonça, A. L. C., and Machado, J. C. (2024). A differentially private guide for graph analytics. In *Proceedings 27th International Conference on Extending Database Technology, EDBT 2024, Paestum, Italy*, pages 850–853.

Camacho, D., Panizo-LLedot, A., Bello-Orgaz, G., Gonzalez-Pardo, A., and Cambria, E. (2020). The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Information Fusion*, 63:88–120.

Chen, L., Han, K., Xiu, Q., and Gao, D. (2022). Graph clustering under weight-differential privacy. In *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 1457–1464. IEEE.

Cormode, G., Procopiuc, C., Srivastava, D., and Tran, T. T. (2012). Differentially private summaries for sparse data. In *Proceedings of the 15th International Conference on Database Theory*, pages 299–311.

Duffield, N., Lund, C., and Thorup, M. (2007). Priority sampling for estimation of arbitrary subset sums. *Journal of the ACM (JACM)*, 54(6):32–es.

Dwork, C. (2006). Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer.

European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. General Data Protection Regulation (GDPR).

Fan, C. and Li, P. (2022). Distances release with differential privacy in tree and grid graph. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2190–2195. IEEE.

Farias, V. A., Brito, F. T., Flynn, C., Machado, J. C., Majumdar, S., and Srivastava, D. (2020). Local dampening: differential privacy for non-numeric queries via local sensitivity. *Proceedings of the VLDB Endowment*, 14(4):521–533.

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.

Federal Communications Commission (2018). Customer privacy. `https://www.fcc.gov/general/customer-privacy`. Online; accessed 13 October 2022.

Ferrara, E., Varol, O., Menczer, F., and Flammini, A. (2016). Detection of promoted social media campaigns. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 563–566.

Ghosh, A., Roughgarden, T., and Sundararajan, M. (2012). Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693.

Hardt, M. and Roth, A. (2012). Beating randomized response on incoherent matrices. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 1255–1268.

Hay, M., Li, C., Miklau, G., and Jensen, D. (2009). Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE 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.

Le Ny, J. and Pappas, G. J. (2013). Privacy-preserving release of aggregate dynamic models. In *Proceedings of the 2nd ACM international conference on High confidence networked systems*, pages 49–56.

Leal, B. C., Vidal, I. C., Brito, F. T., Nobre, J. S., and Machado, J. C. (2018). -doca: Achieving privacy in data streams. In *International Workshop on Data Privacy Management*, pages 279–295. Springer.

Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5–es.

Li, X., Yang, J., Sun, Z., and Zhang, J. (2017). Differential privacy for edge weights in social networks. *Security and Communication Networks*, 2017.

Manríquez, R., Guerrero-Nancuante, C., Martínez, F., and Taramasco, C. (2021). Spread of epidemic disease on edge-weighted graphs from a database: A case study of covid-19. *International Journal of Environmental Research and Public Health*, 18(9):4432.

Matsumoto, H., Yoshida, S., and Muneyasu, M. (2021). Propagation-based fake news detection using graph neural networks with transformer. In *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, pages 19–20. IEEE.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Mendonça, A. L., Brito, F. T., Linhares, L. S., and Machado, J. C. (2017). Dipcoding: a differentially private approach for correlated data with clustering. In *Proceedings of the 21st International Database Engineering & Applications Symposium*, pages 291–297.

Mendonça, A. L., Brito, F. T., and Machado, J. C. (2023). Privacy-preserving techniques for social network analysis. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 174–178. SBC.

Mendonça, A. L., Brito, F. T., and Machado, J. C. (2024). Análise de dados privada em redes sociais. *Jornadas de atualização em informática.*

Monteiro, F. C., Brito, F. T., Chaves, I. C., and Machado, J. C. (2023). Compartilhamento de dados de tráfego de rede utilizando privacidade diferencial. In *Anais do L Seminário Integrado de Software e Hardware*, pages 296–307. SBC.

Neto, E. R., Mendonça, A. L., Brito, F. T., and Machado, J. C. (2018). Privlbs: uma abordagem para preservação de privacidade de dados em serviços baseados em localização. In *Anais do XXXIII Simpósio Brasileiro de Banco de Dados*, pages 109–120. SBC.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Pinot, R., Morvan, A., Yger, F., Gouy-Pailler, C., and Atif, J. (2018). Graph-based clustering under differential privacy. *arXiv preprint arXiv:1803.03831.*

Sealfon, A. (2016). Shortest paths and distances with differential privacy. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 29–41.

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 *Proceedings of the 21st International Database Engineering & Applications Symposium*, pages 74–81.

Varol, O., Ferrara, E., Menczer, F., and Flammini, A. (2017). Early detection of promoted campaigns on social media. *EPJ data science*, 6:1–19.

Wang, D. and Long, S. (2019). Boosting the accuracy of differentially private in weighted social networks. *Multimedia Tools and Applications*, 78(24):34801–34817.