Differentially Private Network Data Collection for Influence Maximization

Extended Abstract

M. Amin Rahimian
University of Pittsburgh
Pittsburgh, PA, USA
rahimian@pitt.edu

Fang-Yi Yu
George Mason University
Fairfax, VA, USA
fangyiyu@gmu.edu

Carlos Hurtado
University of Pittsburgh
Pittsburgh, PA, USA
cah259@pitt.edu

ABSTRACT

When designing interventions in public health, development, and education, decision makers rely on social network data to target a small number of people, capitalizing on peer effects and social contagion to bring about the most welfare benefits to the population. Developing new methods that are privacy-preserving for network data collection and targeted interventions is critical for designing sustainable public health and development interventions on social networks. In a similar vein, social media platforms rely on network data and information from past diffusions to organize their ad campaign and improve the efficacy of targeted advertising. Ensuring that these network operations do not violate users' privacy is critical to the sustainability of social media platforms and their ad economies. We study privacy guarantees for influence maximization algorithms when the social network is unknown, and the inputs are samples of prior influence cascades that are collected at random. Building on recent results that address seeding with costly network information, our privacy-preserving algorithms introduce randomization in the collected data or the algorithm output, and can bound each node's (or group of nodes') privacy loss in deciding whether or not their data should be included in the algorithm input. We provide theoretical guarantees of the seeding performance with a limited sample size subject to differential privacy budgets in both central and local privacy regimes. Simulations on empirical network datasets reveal the diminishing value of network information with decreasing privacy budget in both regimes, as well as additional nuances of post-processing in the local regime.

KEYWORDS

social networks; differential privacy; influence maximization; sample complexity; approximation algorithms

ACM Reference Format:
M. Amin Rahimian, Fang-Yi Yu, and Carlos Hurtado. 2023. Differentially Private Network Data Collection for Influence Maximization: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

When designing interventions in public health, development, and education decision-makers rely on social network data to target a small number of people, capitalizing on peer effects and social contagion to bring about the most welfare benefits to the population [2, 4, 6, 11, 24, 31, 43]. However, members of minority community who contribute to social network data collection endure privacy risks. This is especially important when engaging with vulnerable populations to record sensitive or stigmatizing behavior, e.g., in design of public health interventions among the homeless youth for HIV and suicide prevention [41]. Developing new methods that are privacy-preserving for network data collection and targeted interventions is critical for designing sustainable public health and development interventions on social networks. In a similar vein, social media platforms rely on network data and information from past diffusions to organize their ad campaign and improve the efficacy of targeted advertising. Ensuring that these network operations do not violate users' privacy is critical to the sustainability of social media platforms and their ad economies. This is specially important in light of the vulnerabilities of minority groups in a social network. For example, a few adversarial nodes can coordinate their actions to move a web crawler towards or away from certain parts of an online social network to compromise their privacy through a re-identification attack [3].

Much of the past work on social network data privacy focuses on identifiability of nodes from released graph data, with or without additional information on node attributes and their neighborhoods [35]. Classical anonymization techniques mask node attributes and perturb, modify, randomize or aggregate the graph structure to prevent re-identification of nodes within a confidence level — typically achieving k-anonymity [1, 42, 44]. However, statistical disclosure control of collected data requires safeguarding data donor's privacy against adversarial attacks where common anonymization techniques are shown to be vulnerable to various kinds of identification [5], linkage and cross-referencing [30, 39, 40], statistical difference and re-identification attacks [25]. On the other hand while edge sampling at the data collection can somewhat mitigate the re-identification risk of nodes [17, 35], the seeded nodes can reveal additional side information about who has contributed to network data collection. For example, once a community leader is chosen, his or her close contacts may suffer privacy leaks, even though no social network data are published. Differential privacy (DP)[1] offers an alternative foundation by focusing on protecting input data against revelations of algorithmic outputs [12] with extended applications to optimization problems [15, 29, 36, 38]. The

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

[1] Throughout this paper we use DP to signify any of the phrases "differentially private", "differentially privately", and "differential privacy" with or without hyphen between the words as appropriate by the language function that these words play in the context that DP is used.

influence maximization problem that we address in this paper is a classic example of a cardinality-constrained, combinatorial optimization on a graph with a monotone, submodular objective that admits a tight $(1-1/e)$ approximation guarantee using the greedy node selection algorithm [19–21]. However, the graph structure of the input data has stunned the design of DP seeding algorithms.

2 MAIN RESULTS

We first elaborate on shortcomings of mainstream approaches to DP analysis of graphs for seeding and use that to contextualize our main contributions in proposing new DP definitions with privacy, utility, and sample complexity guarantees for influence maximization.

A privacy notion for social network data collection and intervention design. Influence maximization is a classical, NP-hard network optimization problem, which selects k nodes on a graph of size n to maximize the expected spread size from seeding those k nodes under a randomized model of network diffusion. We focus on differential privacy guarantees for influence maximization when social network is unknown and influential nodes need to be estimated from costly samples of past spreads. The existing literature on DP analyses of graphs, e.g., node- or edge-DP [16, 18, 32], has devoted much attention to graph statistics, e.g., subgraph counts, and is not directly applicable to influence maximization. Furthermore, our hardness results below indicate that common operationalizations of DP for graph algorithms, e.g., node- or edge-DP, are too stringent to render useful performance guarantees for influence maximization. Similar observations have been made about node- and edge-DP guarantees of other graph algorithms [7, 8, 14, 26].

Informal Main Results 1. *Given $n \geq 1$ and $\epsilon > 0$, there can be no ϵ-node-DP or ϵ-edge-DP seeding algorithms that give an approximation guarantee better than $(1-1/e)OPT - \alpha_\epsilon n$ for $\alpha_\epsilon \geq 1/1000e^\epsilon$.*

We formulate new differential privacy notions with guarantees for influence maximization when social network is unknown and the algorithm inputs are samples of prior influence cascades that are collected at random, i.e., *influence samples*, $\mathbf{x} \in \{0,1\}^{n\times m}$, where $\mathbf{x}_{ij}$ indicates appearance of node i in the jth cascade:

Definition 2.1. Given $\epsilon \geq 0$, $n, m \geq 1$, and $k \leq n$, a function $M^k : \{0,1\}^{n\times m} \to \mathcal{Y}$ is *ϵ-influence sample differentially private* if for all outputs $\mathbf{y} \in \mathcal{Y}$ of k seeds, and all pairs of adjacent datasets (collection of influence samples), $\mathbf{x} \sim \mathbf{x}'$ that differ at one entry, we have: $\Pr[M^k(\mathbf{x}) = \mathbf{y}] \leq e^\epsilon \Pr[M^k(\mathbf{x}') = \mathbf{y}]$.

Influence sample differential privacy bounds each node's privacy loss in deciding whether or not their data should be included in the influence samples — as opposed to the social network graph, which is critically different from DP analysis of graphs. Our notion applies when a data donor may be willing to donate most but not all of their contagion history, by providing plausible deniability for a few cascades that can have sensitive or compromising information. In our approach, each node's decision to appear in the input data does not affect the data generation process of the influence samples or the underlying diffusion process; the network graph remains the same and the privacy implications are studied with respect to the construction of the samples on a fixed network. This allows us to meaningfully bound the effect of the removal of a node's data on the algorithmic performance and to use those bounds to tradeoff the privacy risks against improved performance.

Sample complexity with differential privacy and approximation guarantees. Our results build on recent work about seeding with costly network information [13], and more broadly sample complexity of influence maximization [9, 37]. To mitigate privacy risks, we use randomization to provide plausible deniability certificates to those contributing to the input data in one of the two ways: (i) randomizing the algorithm output by exponential mechanisms, i.e., central DP [28], (ii) injecting noise in the input data by randomized response mechanisms, i.e., local DP. Accordingly, we propose two efficient DP seeding algorithms and show their accuracy in terms of the sample size m, network size n, seed set size k, and privacy budget ϵ:

Informal Main Results 2. *Given a graph $\mathcal{G}$ of $n \geq 2$ nodes, and $k \leq n$, for any $0 < \alpha \leq 1$ and $0 < \epsilon < 1/2$, we give a set of algorithms that are centrally or locally ϵ-ISDP and their outputs infect at least $(1-1/e)OPT - \alpha n$ nodes with high probability. Our central ϵ-ISDP algorithm using exponential mechanisms requires at least $m = \max\{\frac{12}{\alpha\epsilon}, \frac{9}{\alpha^2}\}k \ln n$ influence samples and runs in $O(knm)$ time. Our local ϵ-ISDP algorithm uses randomized response mechanism on the influence samples, but requires more (exponentially in k) influence samples: $m = O(k^3\epsilon^{-2k^2} \ln n/\alpha^2)$ with $O(nk^4 + kn^2m)$ run time.*

Exponential mechanism (central DP) only protects the input data against the algorithmic revelations of the output (seeded nodes), giving the algorithm itself unhindered access to user data. On the other hand, the randomized response mechanism (local DP) provides a stronger privacy notion that protects the input data directly against any (adversarial) misuse, regardless of – and including — their use for seeding. Our results address both central and local notions of privacy, providing performance (utility) and sample complexity guarantees for influence maximization algorithms whose outputs/inputs are randomized at a desired level to give central/local DP protection given our input data structure, i.e., influence samples.

3 CONCLUSIONS

The correlated nature of social network information has been noted as a vulnerability in DP formulations which provide user-level guarantees with respect to the addition or removal of a single user's data but ignore information leakage across different users' records [34]. By focusing on the utility of influence samples for influence maximization, our approach guarantees users' participation in any given cascade is kept private either locally or centrally and preserves the statistical validity of the cascades for influence estimation and maximization. Subsequently, we provide a rigorous operationalization of action-level privacy by providing users with plausible deniability as to whether or not they have taken part in any particular network cascade [23]. This relaxation of user-level DP is ideally suited for the purposes of data collection to inform network intervention designs. We expect that our formulation and investigation of influence sample DP, with specific attention to the utility of the collected data for network intervention design, should lead to novel generalizations of privacy that are both consequentialist and robust to variations in sampling distributions and randomization noise [10, 22, 27, 33].

ACKNOWLEDGEMENTS

Rahimian acknowledges support from a Pitt Momentum Funds Grant on Socially Responsible Data Collection and Network Intervention Designs, as well as the computing hardware, software, and research consulting provided through the Pitt Center for Research Computing (Pitt CRC). Rahimian and Yu are listed alphabetically and contributed equally.

REFERENCES

[1] Jemal H Abawajy, Mohd Izuan Hafez Ninggal, and Tutut Herawan. 2016. Privacy preserving social network data publication. *IEEE communications surveys & tutorials* 18, 3 (2016), 1974–1997.

[2] Marcus Alexander, Laura Forastiere, Swati Gupta, and Nicholas A Christakis. 2022. Algorithms for seeding social networks can enhance the adoption of a public health intervention in urban India. *Proceedings of the National Academy of Sciences* 119, 30 (2022), e2120742119.

[3] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. 2007. Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th international conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 181–190.

[4] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2019. Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies* 86, 6 (2019), 2453–2490.

[5] Michael Barbaro, Tom Zeller, and Saul Hansell. 2006. A face is exposed for AOL searcher no. 4417749. *New York Times* 9, 2008 (2006), 8.

[6] Jere R Behrman, Hans-Peter Kohler, and Susan Cotts Watkins. 2002. Social networks and changes in contraceptive use over time: Evidence from a longitudinal study in rural Kenya. *Demography* 39, 4 (2002), 713–738.

[7] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. 2013. Differentially private data analysis of social networks via restricted sensitivity. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. 87–96.

[8] Jeremiah Blocki, Elena Grigorescu, and Tamalika Mukherjee. 2022. Privately Estimating Graph Parameters in Sublinear time. *arXiv preprint arXiv:2202.05776* (2022).

[9] Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier. 2014. Maximizing social influence in nearly optimal time. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 946–957.

[10] Darshan Chakrabarti, Jie Gao, Aditya Saraf, Grant Schoenebeck, and Fang-Yi Yu. 2022. Optimal Local Bayesian Differential Privacy over Markov Chains. (2022). https://doi.org/10.48550/ARXIV.2206.11402

[11] Goylette F. Chami, Sebastian E. Ahnert, Narcis B. Kabatereine, and Edridah M. Tukahebwa. 2017. Social Network Fragmentation and Community Health. *Proceedings of the National Academy of Sciences* 114, 36 (2017), E7425–E7431.

[12] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407.

[13] Dean Eckles, Hossein Esfandiari, Elchanan Mossel, and M. Amin Rahimian. 2022. Seeding with Costly Network Information. *Operations Research* 70, 4 (2022), 2318–2348. https://doi.org/10.1287/opre.2022.2290

[14] Alessandro Epasto, Vahab Mirrokni, Bryan Perozzi, Anton Tsitsulin, and Peilin Zhong. 2022. Differentially Private Graph Learning via Sensitivity-Bounded Personalized PageRank. *arXiv preprint arXiv:2207.06944* (2022).

[15] Anupam Gupta, Katrina Ligett, Frank McSherry, Aaron Roth, and Kunal Talwar. 2010. Differentially private combinatorial optimization. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 1106–1125.

[16] Jacob Imola, Takao Murakami, and Kamalika Chaudhuri. 2021. Locally differentially private analysis of graph statistics. In *30th USENIX Security Symposium (USENIX Security 21)*. 983–1000.

[17] Honglu Jiang, Jian Pei, Dongxiao Yu, Jiguo Yu, Bei Gong, and Xiuzhen Cheng. 2021. Applications of differential privacy in social network analysis: a survey. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[18] Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2013. Analyzing graphs with node differential privacy. In *Theory of Cryptography Conference*. Springer, 457–476.

[19] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 137–146.

[20] David Kempe, Jon Kleinberg, and Éva Tardos. 2005. Influential nodes in a diffusion model for social networks. In *International Colloquium on Automata, Languages, and Programming*. Springer, 1127–1138.

[21] David Kempe, Jon Kleinberg, and Éva Tardos. 2015. Maximizing the Spread of Influence through a Social Network. *Theory of Computing* 11, 4 (2015), 105–147.

[22] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)* 39, 1 (2014), 1–36.

[23] Daniel Kifer, Solomon Messing, Aaron Roth, Abhradeep Thakurta, and Danfeng Zhang. 2020. Guidelines for implementing and auditing differentially private systems. *arXiv preprint arXiv:2002.04049* (2020).

[24] David A Kim, Alison R Hwong, Derek Stafford, D Alex Hughes, A James O'Malley, James H Fowler, and Nicholas A Christakis. 2015. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet* 386, 9989 (2015), 145–153.

[25] Ravi Kumar, Jasmine Novak, Bo Pang, and Andrew Tomkins. 2007. On anonymizing query logs via token-based hashing. In *Proceedings of the 16th international conference on World Wide Web*. 629–638.

[26] George Z Li, Dung Nguyen, and Anil Vullikanti. 2022. Differentially Private Partial Set Cover with Applications to Facility Location. *arXiv preprint arXiv:2207.10240* (2022).

[27] Yuhan Liu, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Michael Riley. 2020. Learning discrete distributions: user vs item-level privacy. *Advances in Neural Information Processing Systems* 33 (2020), 20965–20976.

[28] Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 94–103.

[29] Marko Mitrovic, Mark Bun, Andreas Krause, and Amin Karbasi. 2017. Differentially Private Submodular Maximization: Data Summarization in Disguise. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2478–2487. https://proceedings.mlr.press/v70/mitrovic17a.html

[30] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125.

[31] Elizabeth Levy Paluck, Hana Shepherd, and Peter M Aronow. 2016. Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 566–571.

[32] Sofya Raskhodnikova and Adam Smith. 2016. Lipschitz extensions for node-private graph statistics and the generalized exponential mechanism. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 495–504.

[33] Vibhor Rastogi, Michael Hay, Gerome Miklau, and Dan Suciu. 2009. Relationship privacy: output perturbation for queries with joins. In *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 107–116.

[34] Aria Rezaei and Jie Gao. 2019. On privacy of socially contagious attributes. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1294–1299.

[35] Daniele Romanini, Sune Lehmann, and Mikko Kivelä. 2021. Privacy and uniqueness of neighborhoods in social networks. *Scientific reports* 11, 1 (2021), 1–15.

[36] Omid Sadeghi and Maryam Fazel. 2021. Differentially Private Monotone Submodular Maximization Under Matroid and Knapsack Constraints. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2908–2916.

[37] Gal Sadeh, Edith Cohen, and Haim Kaplan. 2020. Sample Complexity Bounds for Influence Maximization. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[38] Sebastian Perez Salazar and Rachel Cummings. 2021. Differentially Private Online Submodular Maximization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1279–1287.

[39] Latanya Sweeney. 1997. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 25, 2-3 (1997), 98–110.

[40] Latanya Sweeney. 2015. Only you, your doctor, and many others may know. *Technology Science* 2015092903, 9 (2015), 29.

[41] Bryan Wilder, Laura Onasch-Vera, Graham Diguiseppi, Robin Petering, Chyna Hill, Amulya Yadav, Eric Rice, and Milind Tambe. 2021. Clinical Trial of an AI-Augmented Intervention for HIV Prevention in Youth Experiencing Homelessness. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 14948–14956. https://ojs.aaai.org/index.php/AAAI/article/view/17754

[42] Xintao Wu, Xiaowei Ying, Kun Liu, and Lei Chen. 2010. A survey of privacy-preservation of graphs and social networks. In *Managing and mining graph data*. Springer, 421–453.

[43] Amulya Yadav, Bryan Wilder, Eric Rice, Robin Petering, Jaih Craddock, Amanda Yoshioka-Maxwell, Mary Hemler, Laura Onasch-Vera, Milind Tambe, and Darlene Woo. 2017. Influence maximization in the field: The arduous journey from emerging to deployed application. In *Proceedings of the 16th conference on autonomous agents and multiagent systems*. 150–158.

[44] Elena Zheleva and Lise Getoor. 2011. Privacy in social networks: A survey. In *Social network data analytics*. Springer, 277–306.