Privacy and transparency in graph machine learning: A unified perspective

Megha Khosla

Delft University of Technology, Delft, The Netherlands

Abstract

Graph Machine Learning (GraphML), whereby classical machine learning is generalized to irregular graph domains, has enjoyed a recent renaissance, leading to a dizzying array of models and their applications in several domains. With its growing applicability to sensitive domains and regulations by governmental agencies for trustworthy AI systems, researchers have started looking into the issues of transparency and privacy of graph learning. However, these topics have been mainly investigated independently. In this position paper, we provide a unified perspective on the interplay of privacy and transparency in GraphML. In particular, we describe the challenges and possible research directions for a formal investigation of privacy-transparency tradeoffs in GraphML.

Keywords

Graph machine learning, Graph neural networks, Privacy-preserving machine learning, Interpretability/Explainability in machine learning, Post-hoc explainability, Privacy-transparency tradeoffs

1. Introduction

Graphs are a highly informative, flexible, and natural way to represent data. Graph based machine learning (GraphML), whereby classical machine learning is generalized to irregular graph domains, has enjoyed a recent renaissance, leading to a dizzying array of models and their applications in several fields [1, 2, 3, 4, 5]. GraphML models have achieved great success due to their ability to flexibly learn from the complex interplay of graph structure and node attributes/features. Such ability comes with a compromise in privacy and transparency, two indispensable ingredients to achieve trustworthy ML [6].

Deep models trained on graph data are inherently blackbox, and their decisions are difficult for humans to understand and interpret. The growing application of these models in sensitive applications like healthcare and finance and the regulations by various AI governance frameworks necessitate the need for transparency in their decision-making process. Meanwhile, recent research [7, 8, 9, 10] has highlighted the privacy risks of deploying models trained on graph data. It has been suggested that these models are even more vulnerable to privacy leakage than models trained on non-graph data due to the additional encoding of relational structure in the model itself [7].

Consequently, an increasing number of works are focussing on explaining [11, 12, 13, 14] the decisions of black box GraphML models in a post-hoc manner, designing interpretable models [15, 16, 17] as well as privacy preserving techniques for real world deployments of graph models [18, 19, 20].

Despite the growing research interest, the current state of the art considers privacy and transparency in GraphML independently. While transparency provides insight into the model's working, privacy aims to preserve the sensitive information about the training data[1]. The seemingly conflicting goals of privacy and transparency call for the need of a joint investigation. To date, any gain in privacy or transparency is usually compared to any drop in model performance. However, questions like *"what effects would be releasing post-hoc explanations have on the privacy of training data?"* or *"how well can we interpret the decisions of privacy-preserving graph models?"* have so far received little attention [21, 22].

In this position paper, we provide a unified perspective on the inextricable link between privacy and transparency for GraphML. Besides, we sketch the possible research directions towards formally exploring privacy-transparency tradeoffs in GraphML.

2. Background

2.1. Graph Machine Learning

The key idea in graph machine learning is to encode the discrete graph structure into low dimensional continuous vector representations using non-linear dimensionality reduction techniques. Popular classes of GraphML meth-

AIMLAI'22: In Proceedings of Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI) at CIKM'22

m.khosla@tudelft.nl (M. Khosla)
https://khosla.github.io/ (M. Khosla)
0000-0002-0319-3181 (M. Khosla)

[1] Here we are only concerned with data privacy. Model Privacy or protecting the model itself against, for example, stealing model parameters is out of the scope of this paper.

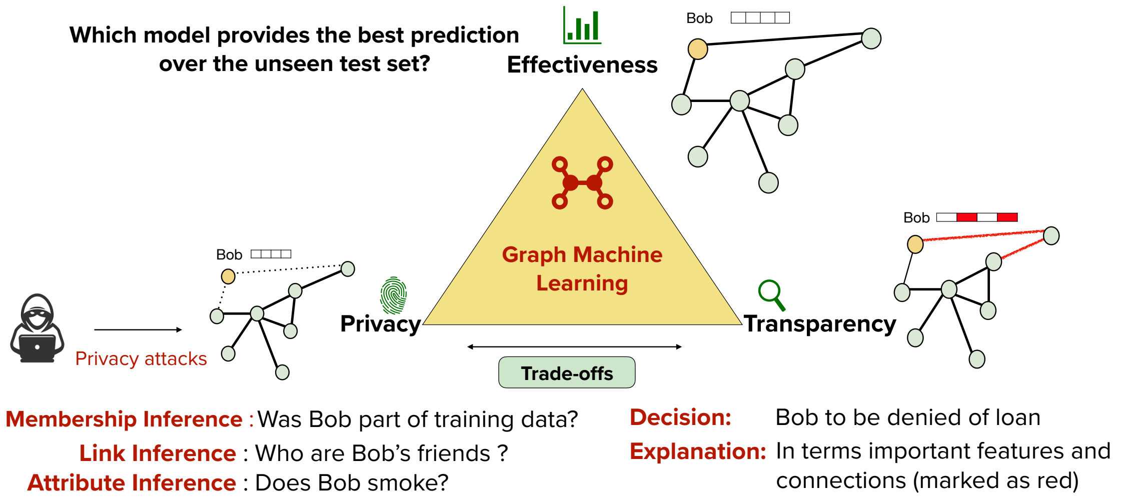

Which model provides the best prediction over the unseen test set?
Effectiveness
Bob
Graph Machine Learning
Bob
Privacy
Transparency
Bob
Privacy attacks
Trade-offs
Membership Inference : Was Bob part of training data?
Link Inference : Who are Bob's friends ?
Attribute Inference : Does Bob smoke?
Decision: Bob to be denied of loan
Explanation: In terms important features and connections (marked as red)

Figure 1: Privacy and transparency are usually studied together with their effect on model performance. But the trade-offs between privacy and transparency have been so far ignored. *Can transparency increase the risk of privacy leakage? How transparent are privacy preserving models?*

ods include *random walk based* strategies [23, 24] which encode structural similarity of the nodes exposed by their co-occurrence in random walks; *matrix-factorization based* [25] which rely on low rank factorization of some node similarity matrix; and the most popular *graph neural networks* (GNNs) [26, 27] which learns node representations by recursive aggregation and transformation of neighborhood features. These methods are usually non-transparent and are shown to be prone to privacy leakage risks.

> Towards improving the adoption of these methods in sensitivity applications like healthcare and medicine the community has started paying attention to the aspects of transparency and privacy. However these aspects have been so far studied independently (see also Figure 1 for an illustration). A formal investigation into the linked role of transparency and privacy in achieving trustworthy GraphML is missing.

2.2. Transparency for GraphML Models

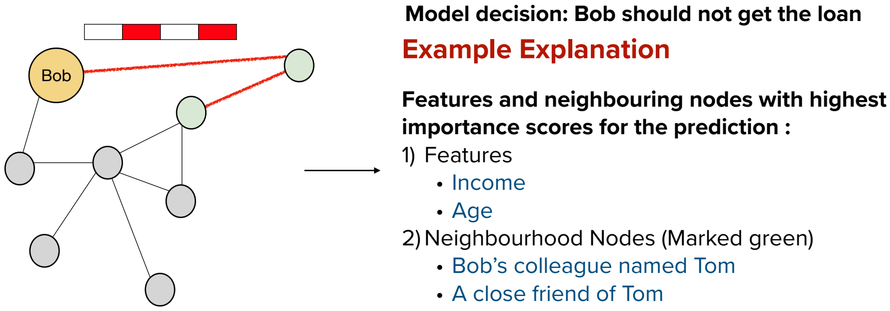
Transparency for deep models, as in GraphML, is usually achieved by providing *explanations* corresponding to decisions of an already trained model or by building *interpretable by design* or *self-explaining* models. Numerous approaches have been proposed in the literature for explaining general machine learning models [28, 29, 30, 31]; however, models learned over graph-structured data have some unique challenges.

Specifically, predictions on graphs are induced by a complex combination of nodes and paths of edges between them in addition to the node features. A trivial application of existing explainability methods to graph models cannot account for the role of graph structure in the model decision. Consequently several graph specific explainability approaches have been recently developed which focus primarily on explaining graph neural networks' decisions for node and graph classification [32, 33].

Explanations usually include the importance scores for nodes/edges in a subgraph (or node's neighborhood in case of node-level task) and the node features [11, 12, 13]. Figure 2 depicts an example of an explanation over graph data. Depending on the explanation method, the importance scores could be either continuous (soft masks) or binary (hard masks). A few works have also been proposed to explain dense unsupervised node representations [34, 35]. In terms of methodologies, several techniques based on input perturbations [11, 12, 13], input gradients[36, 37], causal techniques [34, 38, 33] as well as utilizing simpler surrogate models [14] have been explored.

Another methodology to provide transparency is to develop interpretable by design models [15, 16, 39]. Such models usually contain a self-explanatory module trained jointly with the learner module. Explanations are thus, by design, faithful to the model.

A few other works also focus on unifying diverse notions of evaluation strategies [40, 37] necessary for effectively assessing the quality and utility of explanations.



Model decision: Bob should not get the loan

Example Explanation

Features and neighbouring nodes with highest importance scores for the prediction :

1) Features
   - Income
   - Age
2) Neighbourhood Nodes (Marked green)
   - Bob's colleague named Tom
   - A close friend of Tom

Figure 2: An example explanation in terms of features and node attribution over a social network in which a node represents a user and edges represent friendship relation. Node features correspond to demographic attributes of the user. Neighboring nodes with high importance scores are marked green.

> Despite the progress in improving transparency of GraphML techniques its effect on data privacy has escaped attention. While transparency could increase the utility of the model, for sensitive applications any unaddressed concerns for privacy can hinder the full adoption of the models and further dissuade the participants to share their data.

2.3. Privacy in GraphML

Deep learning models, in general, are known to leak private information about the employed training data. Recent works showed that trained model on graph data can leak sensitive information about the training data (see Figure 3) like node membership [7, 8], certain dataset properties [41] and connectivity structure of the nodes [9]. In Figure 3 we illustrate the possibility of different privacy attacks given access to trained GraphML model. Compared to general deep learning models, GraphML is more vulnerable to privacy risks as they incorporate not only the node features/labels but also the graph structure [7].

Privacy-preserving techniques for graph models are mainly based on differential privacy [42, 7, 19, 20] and adversarial training frameworks [43, 44, 45]. The key idea in differential privacy [46] is to conceal the presence of a single individual in the dataset. In particular, if we query a dataset containing N individuals, the query's result will be probabilistically indistinguishable from the result of querying a neighboring dataset with one less or one more individual. For machine learning models, such probabilistic indistuinguishability is achieved by adding appropriate levels of noise at different levels of model development. For instance, [42] employs objective perturbation mechanism to develop differential private network embeddings. Olatunji et al. [7] combines the knowledge-distillation framework with the two noise mechanisms, random subsampling, and noisy labeling to release graph neural networks under differential privacy guarantees. In particular it uses only a random sample of private data to train teacher models corresponding to nodes in an unlabelled public dataset. The final model which is later released is trained on public data using the noisy labels generated by the teacher models. Other works [20, 19] do not build a separate public model but achieve DP via adding noise directly to the aggregation module of GNNs. Adversarial defence to privacy attacks on GNNs is proposed in [43], in which the predictability of private labels is destroyed and the utility of perturbed graphs is maintained. An adversarial learning approach based on mini-max game between the desired graph feature encoder and the worst-case attacker is proposed in [44] to address the attribute inference attack on GNNs.

> Despite the growing number of works in improving privacy in GraphML, its effect on transparency of these models is not at all studied. The complex mechanisms employed to ensure privacy further hurts the model transparency. Consequently it is not clear if existing explainers can be used to explain the decision making process of privacy-preserving models.

3. A Unified Perspective

Graphs are powerful abstractions that facilitate leveraging data interconnection to represent, predict, and explain real-world phenomena. Exploiting such explicit or latent data interconnections, on the one hand, makes GraphML more powerful but also brings in additional challenges, further exacerbating the need for a joint investigation of privacy and transparency. In following

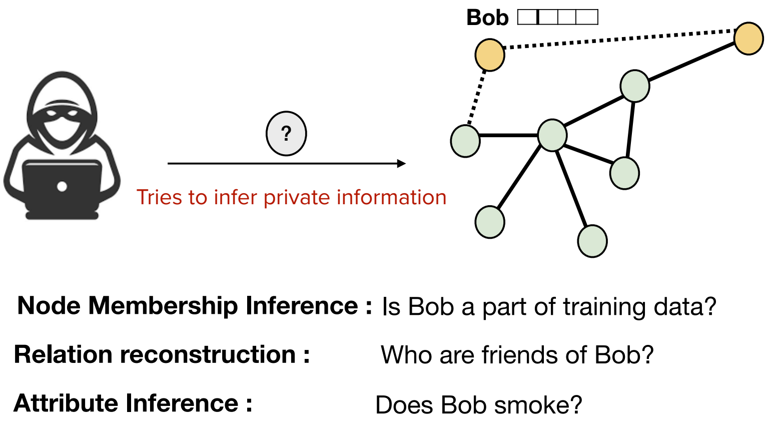

Bob
?
Tries to infer private information
Node Membership Inference : Is Bob a part of training data?
Relation reconstruction : Who are friends of Bob?
Attribute Inference : Does Bob smoke?

Figure 3: Given access to trained model or embeddings trained on graph data, an adversary can launch several attacks to infer membership, relations or attributes of a node.

we discuss the key issues arising due to the independent treatment of privacy and transparency for GraphML.

3.1. Diverse explanation types and methods

Model explanations for graph data are usually in the form of feature and neighborhood (subgraph) attributions. In particular, importance scores for node features and its neighboring nodes/edges are released as explanations. Neighborhood attributions or structure explanations are a more direct form of information leakage. They can be, for example, leveraged to identify nodes in the training set or infer hidden attributes of sensitive nodes using the attributes of their neighbors.

Besides, the data points (nodes) in graph data are correlated, thus violating the usual i.i.d. assumption over data distributions. Consequently, the decisions and explanations over correlated nodes might themselves be correlated. Such correlations among released explanations might be exploited to reconstruct sensitive information of the training data. For example, the similarity in feature explanations for recommendations to two connected users might reveal the sensitive link information they want to hide. Towards this [22] show that the link structure of the training graph can be reconstructed with a high success rate even if only the feature explanations are available.

3.2. Transparency of private models

Moreover, due to the correlated nature of the graph data, privacy-preserving mechanisms on graph models need to focus on several aspects such as node privacy, edge privacy, and attribute privacy [20]. This leads to more complex privacy-preserving mechanisms, which results in a further loss of transparency. To understand the issue, consider a simple differential privacy-based mechanism in which randomized noise is added to the model's output. Such noise could alter the final decision but not the decision process that an explanation (according to its current definition) is usually expected to reveal. Model agnostic approaches for explainability, which only assume black-box access to the trained model, might be misguided by such alteration in the final decision.

3.3. The curse of overfitting

In traditional machine learning, we can randomly divide the data into two parts to obtain training and test sets. It is more tricky in graphs where the data points are connected, and random data sampling may result in non i.i.d. train and test sets. Even for the task of graph classification where the graphs constitute the datapoints instead of the the nodes, distributional changes are common in train and test splits [47] due to varying graph structure and size. Specifically, the train set may contain specific spurious correlations which are not representative of the entire dataset. This puts GraphML models at a higher risk of overfitting to sample specific correlations rather than learning the desired general patterns [48]. Existing privacy attacks have leveraged overfitting to reveal sensitive information about the training sample [49]. Exploiting associated explanations, which in principle should reveal learned spurious correlations, can further aid in privacy leakage.

4. Research Directions

Based on the described issues and challenges in the previous section, we recommend the following research directions towards a formal investigation of privacy-transparency tradeoffs.

1. **New Threat Models.** A first step is to quantify the privacy risks of releasing post-hoc explanations. Towards that, we need to design new threat models and *structure-aware* privacy attacks in the presence of post-hoc model explanations. Care should be taken to formulate *realistic assumptions on adversary's background knowledge.* For example, in highly homophilic graphs, an adversary might be able to approximate well the link structure of the graph only if the node features/labels are available. *What information explanations could leak in addition when explanations are provided?*

2. **Risk-utilty assessment of different explanation types and methods.** Model explanations for GraphML can be in the form of feature or node/edge importance scores. Besides, existing explanation methods are based on different methodologies and might be discovering different aspects of the model decision process. Depending on the dataset and application, certain types of explanation methods and types of explanation (feature or structural) might be preferred over others. A dataset and application-specific risk-utility assessment might reveal more favorable explanations for minimizing privacy loss. For instance, [22] finds that gradient-based feature explanations have the least predictive (faithfulness to the model) power for the task of node classification but leak the most amount of information about the private structure of the training graph. In such cases, one can decide not to reveal such an explanation as it has little utility for the user.

3. **Transparency of privacy-preserving models.** Besides evaluating the privacy risks of releasing explanations, it is essential to analyze the transparency of privacy-preserving techniques. It is not clear if existing explanation strategies can faithfully explain the privacy-preserving models' decisions. Questions like *what should be the properties of explanations of such models? What constitutes a faithful explanation?* need to be investigated. Consequently new techniques to explain privacy preserving models need to be developed.

4. **Reducing overfitting.** Overfitting is usually considered a common enemy for model effectiveness on unseen data and privacy. Recently, a few works have proposed interpretable by design models for example using stochastic attention mechanisms [39], graph sparsification strategies [16] etc. These methods are claimed to remove spurious correlations in the training phase leading to a reduction in overfitting. A possible research direction is further exploiting such transparency strategies to minimize privacy leakage.

5. Conclusion

There has been an unprecedented rise in the popularity of graph machine learning in recent years. With its growing applications in sensitive areas, several works focus independently on their transparency and privacy aspects. We provide a unified perspective on the need for a joint investigation of privacy and transparency in GraphML. We hope to start a discussion and foster future research in quantifying and resolving the privacy-transparency tradeoffs in GraphML. Resolution of such tradeoffs would make GraphML more accessible to stakeholders currently tied down by regulatory concerns and lack of trust in the solutions.

References

[1] T. Gaudelet, B. Day, A. R. Jamasb, J. Soman, C. Regep, G. Liu, J. B. R. Hayter, R. Vickers, C. Roberts, J. Tang, D. Roblin, T. L. Blundell, M. M. Bronstein, J. P. Taylor-King, Utilizing graph machine learning within drug discovery and development, Briefings in Bioinformatics 22 (2021). doi:10.1093/bib/bbab159.

[2] T. N. Dong, S. Mucke, M. Khosla, Mucomid: A multitask graph convolutional learning framework for mirna-disease association prediction, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2022).

[3] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, J. Leskovec, Graph convolutional neural networks for web-scale recommender systems, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '18, ACM, 2018, pp. 974–983.

[4] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, P. Battaglia, Learning to simulate complex physics with graph networks, in: International Conference on Machine Learning, PMLR, 2020, pp. 8459–8468.

[5] T. N. Dong, S. Johanna, S. Mucke, M. Khosla, A message passing framework with multiple data integration for mirna-disease association prediction, Scientific Reports (2022). doi:10.1038/s41598-022-20529-5.

[6] E. Dai, T. Zhao, H. Zhu, J. Xu, Z. Guo, H. Liu, J. Tang, S. Wang, A comprehensive survey on trust-

worthy graph neural networks: Privacy, robustness, fairness, and explainability, arXiv preprint arXiv:2204.08570 (2022).

[7] I. E. Olatunji, W. Nejdl, M. Khosla, Membership inference attack on graph neural networks, in: 2021 IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 11–20.

[8] V. Duddu, A. Boutet, V. Shejwalkar, Quantifying privacy leakage in graph embedding, in: MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2020, pp. 76–85.

[9] Z. Zhang, Q. Liu, Z. Huang, H. Wang, C. Lu, C. Liu, E. Chen, Graphmi: Extracting private graph data from graph neural networks, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, 2021, pp. 3749–3755.

[10] X. He, J. Jia, M. Backes, N. Z. Gong, Y. Zhang, Stealing links from graph neural networks, in: 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 2669–2686.

[11] R. Ying, D. Bourgeois, J. You, et al., GNN explainer: A tool for post-hoc explanation of graph neural networks, Advances in neural information processing systems 32 (2019) 9240–9251.

[12] T. Funke, M. Khosla, M. Rathee, A. Anand, Zorro: Valid, sparse, and stable explanations in graph neural networks, IEEE Transactions on Knowledge and Data Engineering (2022) 1–12. doi:10.1109/TKDE.2022.3201170.

[13] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, X. Zhang, Parameterized explainer for graph neural network, Advances in Neural Information Processing Systems 33 (2020).

[14] M. N. Vu, M. T. Thai, Pgm-explainer: Probabilistic graphical model explanations for graph neural networks, in: NeurIPS, 2020.

[15] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, R. He, Graph information bottleneck for subgraph recognition, arXiv preprint arXiv:2010.05563 (2020).

[16] M. Rathee, Z. Zhang, T. Funke, M. Khosla, A. Anand, Learnt sparsification for interpretable graph neural networks, arXiv preprint arXiv:2106.12920 (2021).

[17] Z. Zhang, Q. Liu, H. Wang, C. Lu, C. Lee, Protgnn: Towards self-explaining graph neural networks, arXiv preprint arXiv:2112.00911 (2021).

[18] I. E. Olatunji, T. Funke, M. Khosla, Releasing graph neural networks with differential privacy guarantees, arXiv preprint arXiv:2109.08907 (2021).

[19] S. Sajadmanesh, D. Gatica-Perez, Locally private graph neural networks, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 2130–2145.

[20] S. Sajadmanesh, A. S. Shamsabadi, A. Bellet, D. Gatica-Perez, Gap: Differentially private graph neural networks with aggregation perturbation, arXiv preprint arXiv:2203.00949 (2022).

[21] R. Shokri, M. Strobel, Y. Zick, On the privacy risks of model explanations, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 231–241. URL: https://doi.org/10.1145/3461702.3462533. doi:10.1145/3461702.3462533.

[22] I. E. Olatunji, M. Rathee, T. Funke, M. Khosla, Private graph extraction via feature explanations, in: Accepted for publication in 23rd Privacy Enhancing Technologies Symposium (PETS 2023), 2023. URL: https://arxiv.org/abs/2206.14724.

[23] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: KDD, 2014.

[24] M. Khosla, J. Leonhardt, W. Nejdl, A. Anand, Node representation learning for directed graphs, in: ECML, 2019.

[25] C. Zhou, Y. Liu, X. Liu, Z. Liu, J. Gao, Scalable graph embedding for asymmetric proximity., in: AAAI, 2017, pp. 2942–2948.

[26] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: ICLR, 2017.

[27] W. L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, in: NeurIPS, 2017.

[28] J. Chen, L. Song, M. J. Wainwright, M. I. Jordan, Learning to explain: An information-theoretic perspective on model interpretation, arXiv:1802.07814 (2018).

[29] J. Yoon, J. Jordon, M. van der Schaar, Invase: Instance-wise variable selection using neural networks, ICLR (2018).

[30] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, W. Samek, Layer-wise relevance propagation for neural networks with local renormalization layers, in: ICANN, 2016.

[31] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: PMLR, 2017.

[32] H. Yuan, J. Tang, X. Hu, S. Ji, Xgnn: Towards model-level explanations of graph neural networks, in: SIGKDD, 2020.

[33] Y. Gao, T. Sun, R. Bhatt, D. Yu, S. Hong, L. Zhao, Gnes: Learning to explain graph neural networks, in: ICDM, 2021.

[34] B. Kang, J. Lijffijt, T. De Bie, Explaine: An approach for explaining network embedding-based link predictions, arXiv:1904.12694 (2019).

[35] M. Idahl, M. Khosla, A. Anand, Finding interpretable concept spaces in node embeddings using

knowledge bases, in: Workshops of ECML PKDD, 2019.

[36] P. E. Pope, S. Kolouri, M. Rostami, et al., Explainability methods for graph convolutional neural networks, in: CVPR, 2019.

[37] B. Sanchez-Lengeling, J. Wei, B. Lee, E. Reif, P. Wang, W. W. Qian, K. McCloskey, L. Colwell, A. Wiltschko, Evaluating attribution for graph neural networks, NeurIPS (2020).

[38] M. Bajaj, L. Chu, Z. Y. Xue, J. Pei, L. Wang, P. C.-H. Lam, Y. Zhang, Robust counterfactual explanations on graph neural networks, Advances in Neural Information Processing Systems 34 (2021) 5644–5655.

[39] S. Miao, M. Liu, P. Li, Interpretable and generalizable graph learning via stochastic attention mechanism, arXiv preprint arXiv:2201.12987 (2022).

[40] M. Rathee, T. Funke, A. Anand, M. Khosla, Bagel: A benchmark for assessing graph neural network explanations, 2022. URL: https://arxiv.org/abs/2206.13983. doi:10.48550/ARXIV.2206.13983.

[41] Z. Zhang, M. Chen, M. Backes, Y. Shen, Y. Zhang, Inference attacks against graph neural networks, in: Proc. USENIX Security, 2022.

[42] D. Xu, S. Yuan, X. Wu, H. Phan, Dpne: Differentially private network embedding, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 235–246.

[43] I.-C. Hsieh, C.-T. Li, Netfense: Adversarial defenses against privacy attacks on neural networks for graph data, IEEE Transactions on Knowledge and Data Engineering (2021) 1–1. doi:10.1109/TKDE.2021.3087515.

[44] P. Liao, H. Zhao, K. Xu, T. Jaakkola, G. J. Gordon, S. Jegelka, R. Salakhutdinov, Information obfuscation of graph neural networks, in: International Conference on Machine Learning, PMLR, 2021, pp. 6600–6610.

[45] K. Li, G. Luo, Y. Ye, W. Li, S. Ji, Z. Cai, Adversarial privacy-preserving graph embedding against inference attack, IEEE Internet of Things Journal 8 (2020) 6904–6915.

[46] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of cryptography conference, Springer, 2006, pp. 265–284.

[47] H. Li, X. Wang, Z. Zhang, W. Zhu, Out-of-distribution generalization on graphs: A survey, arXiv preprint arXiv:2202.07987 (2022).

[48] Q. Zhu, N. Ponomareva, J. Han, B. Perozzi, Shift-robust gnns: Overcoming the limitations of localized graph training data, Advances in Neural Information Processing Systems 34 (2021) 27965–27977.

[49] S. Yeom, I. Giacomelli, A. Menaged, M. Fredrikson, S. Jha, Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning, Journal of Computer Security 28 (2020) 35–70.