A Review of Poisoning Attacks on Graph Neural Networks

Kihyun Seol, Yerin Lee, Seungyeop Song, Heejae Park, and Laihyuk Park
Department of Computer Science and Engineering, Seoul National University of Science and Technology, Seoul, 01811, Korea
Email: {seolpark, 21101146, thdtmdduqdhk, prkhj98, lhpark}@seoultech.ac.kr

***Abstract*—A Graph Neural Network (GNN) is designed to generate effective node embeddings in graph-structured data. Therefore, GNNs are well-suited for tasks like node classification and graph generation. As their use has expanded, concerns about their security, robustness, and privacy have grown. This paper explores the various adversarial attacks that can be executed against GNNs, focusing on poisoning attacks, a specific type of adversarial attack. We examine key attack strategies and review recent research developments in each attack. Finally, we conclude with proposals aimed at enhancing the robustness of GNNs.**

***Index Terms*—Adversarial attack, poisoning attack, GNNs.**

I. Introduction

Graph Neural Networks (GNNs) are highly effective at generating network representations based on message-passing algorithms. They excel in representing unstructured network data, making them particularly suitable for tasks such as node classification, link prediction, clustering, graph generation, and physical simulation [1], [2]. Additionally, GNNs possess remarkable generalization capabilities, enabling them to effectively manage unseen data during training. This strength has facilitated their extensive application, extending beyond network-related domains to a wide array of academic disciplines. In this context, concerns about the security of GNN models in these various domains have arisen, with increasing emphasis on the need for robustness and privacy [3]. Adversarial attacks, originally developed to evaluate the robustness of traditional deep neural networks, can similarly be applied to GNNs, contributing to the development of trustworthy AI systems. In this paper, we provide an overview of adversarial attacks applicable to GNNs, with a particular focus on poisoning attacks, which are considered the most practical adversarial attacks in the context of GNNs [4].

II. Preliminaries of Adversarial Attacks

GNNs are vulnerable to adversarial attacks. The impact of an adversarial attack can vary depending on the attacker's knowledge of the dataset and the trained model. The attacks can be categorized into three levels, i.e., white-box, gray-box, and black-box. While white-box attacks might appear unrealistic because they involve full access to model parameters, the adjacency matrix, and labels, they are commonly used to demonstrate the worst-case scenario under adversarial attacks. The gray-box attack involves an attacker with framework access similar to a white-box attack but with limited knowledge, leading to less impact than white-box attacks. In a black-box attack, the attacker does not have access to the model parameters but can still interact with parts of the graph dataset, such as graph structure or specific nodes, and modify these features.

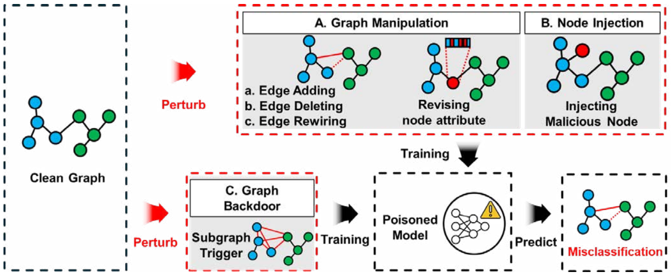

Fig. 1. Three types of poisoning attacks on GNNs.

While there are various types of adversarial attacks, one that compromises integrity is the poisoning attack. Poisoning attack stands out for their ability to inject a small amount of malicious data (such as node) in the training phase, which can significantly impair the model's effectiveness. This attack method is especially applicable to graph mining tasks based on transductive learning, where the goal is to make predictions only on a specific set of test instances known at the time of training. As a result, Poisoning attack is particularly well-suited for semi-supervised node classification tasks.

III. Poisoning Attacks on Graph Neural Networks

Representative adversarial attacks resulting from poisoning attacks include graph manipulation attacks, node injection attacks, and backdoor attacks, as shown in Fig 1.

A. Graph Manipulation

Recently, many studies on poisoning-based adversarial attacks have focused on manipulating the graph structure. These approaches often involve adding/deleting specific edges, or revising node attributes. Many existing types of research have altered the graph structure by adding or deleting edges and modifying node attributes to induce errors in node labeling tasks [5], [6]. However, the study [7] introduced a different approach by applying a deep reinforcement learning-based rewiring operation. In this rewiring operation, the graph retains the same number of edges and nodes, and its total degree

979-8-3503-6463-7/24/$31.00 ©2024 IEEE

remains constant. The operation makes only minor adjustments to the first few eigenvalues of the graph's adjacency matrix, making the adversarial perturbations unnoticeable.

B. Node Injection

Node injection is performed by adding only malicious nodes without affecting the existing edges or node attributes. The authors in [8] proposed Node Injection Poisoning Attacks (NIPA), which involve adding fake nodes. They use a hierarchical reinforcement learning approach for the objective function. This reinforcement learning approach is inspired by RL-S2V [9], which was proposed for evasion attacks and has been further developed into a hierarchical Q-network to enhance efficiency in the search process. In addition, ongoing research aims to ensure that attacks can be effectively executed even under challenging conditions. For instance, the authors in [10] proposed Graph-Attack Advanced Actor-Critic (GA2C), a method for performing node injection in a black-box setting with restricted graph information. This approach demonstrated superior results using the advantage actor-critic algorithm. In the same context, the Generalizable Node Injection Attack model (G-NIA) [11] focused on scalability by performing node injection with just a single node, drastically reducing the cost of the attack. Recently, the authors in [12] proposed Node Injection for Class-specific Network Poisoning (NICKI) that implemented a two-phase learning process: the first phase learns the node representation, while the second phase generates the features and edges of the injected nodes. It was observed that unlike previous baseline models such as NIPA [8] and Approximate Fast Gradient Sign Method (AFGSM) [13], where performance did not improve with larger budgets for edge and feature perturbations, NICKI showed improved attack performance as the budget increased.

C. Graph Backdoor

A graph backdoor attack is one of the most recently studied areas among the three types of poisoning attacks. This attack is executed by injecting a backdoor trigger into the training set, which can be either an attribute of a single node or a subgraph that aligns with a predefined pattern. In a recent study, the authors in [14] proposed a Semantic Backdoor Attack against GCNs (SBAG), which is a black-box semantic backdoor attack on GCNs. In SBAG, a specific type of node is used as the backdoor trigger, and when this backdoor is activated, the GCN misclassifies the target. Specifically, SBAG inserts the backdoor through a two-step process right before training: in the first step, it selects a semantic trigger node, and in the second step, it generates poisoning samples by selecting samples with the top-k scores. The study observed that the backdoor was successfully activated with high probability across four datasets—AIDS, NCI1, PROTEINS, and ENZYMES—and maintained a high success rate even when the poisoning rate was as low as 5%.

IV. Conclusion

In this paper, we explored key poisoning attack strategies, including graph manipulation, node injection, and graph backdoor attacks, while reviewing current research trends. Nowadays, in response to such attacks, researchers are developing trustworthy AI across various industries. However, deploying trustworthy GNNs in real-world applications remains challenging. Notably, adversarial attacks may be leveraged during the pre-training of GNNs, leading to the incorporation of graph backdoor attacks following fine-tuning. Furthermore, as adversarial attacks increasingly attempt to perturb models by introducing label noise, ongoing research is required to develop GNNs robust to label noise for real-world applications.

Acknowledgment

This research was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2022-00156353) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and IITP grant funded by the Korea government(MSIT) (No.RS-2024-00437252, Development of anti-sniffing technology in mobile communication and AirGap environments).

References

[1] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 249–270, 2020.

[2] G. Li, T.-H. Nguyen, and J. J. Jung, "Traffic incident detection based on dynamic graph embedding in vehicular edge computing," *Applied Sciences*, vol. 11, no. 13, p. 5861, 2021.

[3] L. Sun, Y. Dou, C. Yang, K. Zhang, J. Wang, S. Y. Philip, L. He, and B. Li, "Adversarial attack and defense on graph data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 7693–7711, 2022.

[4] E. Dai, T. Zhao, H. Zhu, J. Xu, Z. Guo, H. Liu, J. Tang, and S. Wang, "A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability," *arXiv preprint arXiv:2204.08570*, 2022.

[5] A. Bojchevski and S. Günnemann, "Adversarial attacks on node embeddings via graph poisoning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 695–704.

[6] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 2847–2856.

[7] Y. Ma, S. Wang, T. Derr, L. Wu, and J. Tang, "Graph adversarial attack via rewiring," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1161–1169.

[8] Y. Sun, S. Wang, X. Tang, T.-Y. Hsieh, and V. Honavar, "Adversarial attacks on graph neural networks via node injections: A hierarchical reinforcement learning approach," in *Proceedings of the Web Conference 2020*, 2020, pp. 673–683.

[9] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song, "Adversarial attack on graph structured data," in *International conference on machine learning*. PMLR, 2018, pp. 1115–1124.

[10] M. Ju, Y. Fan, Y. Ye, and L. Zhao, "Black-box node injection attack for graph neural networks," *arXiv preprint arXiv:2202.09389*, 2022.

[11] S. Tao, Q. Cao, H. Shen, J. Huang, Y. Wu, and X. Cheng, "Single node injection attack against graph neural networks," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1794–1803.

[12] A. K. Sharma, R. Kukreja, M. Kharbanda, and T. Chakraborty, "Node injection for class-specific network poisoning," *Neural Networks*, vol. 166, pp. 236–247, 2023.

[13] J. Wang, M. Luo, F. Suya, J. Li, Z. Yang, and Q. Zheng, "Scalable attack on graph data by injecting vicious nodes," *Data Mining and Knowledge Discovery*, vol. 34, pp. 1363–1389, 2020.

[14] J. Dai, Z. Xiong, and C. Cao, "A semantic backdoor attack against graph convolutional networks," *Neurocomputing*, vol. 600, p. 128133, 2024.