Poster: Multi-target & Multi-trigger Backdoor Attacks on Graph Neural Networks

Jing Xu
Delft University of Technology
Delft, Netherlands
j.xu-8@tudelft.nl

Stjepan Picek
Radboud University & Delft University of Technology
Nijmegen, Netherlands
stjepan.picek@ru.nl

ABSTRACT

Recent research has indicated that Graph Neural Networks (GNNs) are vulnerable to backdoor attacks, and existing studies focus on the One-to-One attack where there is a single target triggered by a single backdoor. In this work, we explore two advanced backdoor attacks, i.e., the multi-target and multi-trigger backdoor attacks, on GNNs: 1) One-to-N attack, where there are multiple backdoor targets triggered by controlling different values of the trigger; 2) N-to-One attack, where the attack is only triggered when all the N triggers are present. The initial experimental results illustrate that both attacks can achieve a high attack success rate (up to 99.72%) on GNNs for the node classification task.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

Backdoor Attacks; Graph Neural Networks; Node Classification

ACM Reference Format:
Jing Xu and Stjepan Picek. 2023. Poster: Multi-target & Multi-trigger Backdoor Attacks on Graph Neural Networks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3576915.3624387

1 INTRODUCTION

With the increasing development of GNNs in security applications [2, 10], studying the backdoor attacks on GNNs is crucial. In the existing studies about backdoor attacks on GNNs, the attacker modifies the training dataset by injecting a backdoor trigger into some training samples and relabeling these samples to a predetermined target label. Then, the backdoored GNN model, which is trained on the backdoored training dataset, will output the attacker-chosen label when a trigger is injected into a testing sample.

So far, most existing research on backdoor attacks on CNNs focuses on attacking a single target class and being triggered by a single backdoor trigger, which can be defined as *One-to-One* attack. In addition to the *One-to-One* attack, it has been demonstrated for CNNs that multi-target and multi-trigger backdoor attacks can achieve more stealthy and powerful attack performance [9]. Specifically, the multi-target backdoor attacks on CNNs can be implemented by controlling the different intensities of the same backdoor trigger to trigger multiple backdoor target labels, which can be defined as *One-to-N* attack. In the multi-trigger backdoor attacks on CNNs, the backdoor attack can only be triggered when all backdoor triggers are satisfied, while any single backdoor trigger cannot trigger the backdoor attack, which can be defined as *N-to-One* attack. However, existing research on backdoor attacks on GNNs focuses on attacking a single target class. These attacks are only triggered by a single backdoor trigger, which is *One-to-One* attack.[1]. Although multi-target and multi-trigger backdoor attacks have been proposed in the image domain [9], to the best of our knowledge, there is no work on the multi-target and multi-trigger backdoor attacks on GNNs. We here aim to bridge this gap. Our preliminary results show that both *One-to-N* and *N-to-One* backdoor attacks can achieve high attack success rates on GNNs.

2 METHODOLOGY

2.1 Problem Formulation

GNNs take a graph $G = (V, E, X)$ as an input, where V, E, X denote nodes, edges, and node attributes, and learn a representation vector (embedding) for each node, z_v ($v \in G$), or the entire graph, z_G. Modern GNNs update the representation of a node by aggregating representations of its neighbors. After k iterations of aggregation, a node's representation captures structure and feature information within its k-hop network neighborhood. For the node classification task, the node representation z_v is used for prediction. The node classification task aims to predict the class label(s) for the unlabeled nodes using the node representations z_v. We consider a *gray-box* threat model, which assumes that the attacker is able to freely modify a small portion of the training dataset. We also assume the attacker performs a *dirty-label* backdoor attack, changing the labels of the poisoned samples to the target label(s). Although this kind of attack is weaker than *clean-label* backdoor attacks [4], where the labels remain unaltered, dirty label attacks are the most common in the literature [6, 8, 11]. The goal of the attacker is to inject a backdoor in the given pre-trained clean GNN model through training over the poisoned training dataset, which achieves misclassification under the presence of a trigger while maintaining clean high accuracy on the original task. This threat model is realistic in real-world settings. For example, in cases where the training dataset is collected from public users, the adversary can provide

[1] In the Federated GNNs, there is a work considering multiple backdoor triggers [7], but here we focus on centralized GNNs.

CCS '23, November 26–30, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0050-7/23/11.
https://doi.org/10.1145/3576915.3624387

trigger-embedded training data to implement the backdoor attack. We focus on utilizing the feature-trigger backdoor attack from [8] for multi-target and multi-trigger backdoor attacks on GNNs for the node classification task. The trigger used in the backdoor attacks in our paper is defined as follows, as well as the One-to-N and N-to-One attacks:

Definition 2.1 (Trigger). A specific feature pattern created by modifying the value of a subset of a node's features.

Definition 2.2 (One-to-N Attack). The adversary can trigger multiple backdoor targets by controlling the different values of the same feature trigger on the target node.

Definition 2.3 (N-to-One Attack). Multiple feature triggers are in different locations in the feature vector and have the same value. The adversary can trigger the backdoor attack only when all the feature triggers are satisfied, while any single backdoor trigger cannot trigger the backdoor.

2.2 General Framework

In our work, both the multi-target and multi-trigger backdoor attacks include two phases: generating backdoored training datasets and embedding the backdoor into the pretrained GNN model Φ_o. In the backdoored datasets generating phase, the adversary samples data from the original training dataset, which are not from the target classes. For each selected training sample, the adversary injects specific trigger(s) and changes their label into the target classes to generate the backdoored training datasets. Then, after training the pre-trained GNN model Φ_o with the backdoored training dataset, the backdoor is embedded into the GNN model resulting in the backdoored GNN model Φ_b. Specifically, we demonstrate the general framework of the *One-to-N* and *N-to-One* attacks as follows.

2.2.1 One-to-N Attack. The One-to-N attack is able to trigger N different backdoor targets $(y_{t_1}, y_{t_2}, \cdots y_{t_N})$ by controlling different feature values $(v_1, v_2, \cdots, v_N)$ of the same feature trigger. If the feature trigger of value v_i is injected into the target node, the backdoored model Φ_b will misclassify the target node as the corresponding target class y_{t_i}. The attack framework of the One-to-N attack is presented in Figure 1 and we assume there are 2 backdoor targets in the One-to-N attack, i.e., $N = 2$. In the training phase, the attacker samples data from the original training dataset, which is in the non-target classes. Then, the selected training samples are injected into a feature trigger t with value v_1 or v_2. For all the selected training samples, the trigger injecting position of the feature trigger is the same, e.g., in the last three dimensions of the feature vector of the node as shown in Figure 1, while the feature values of the feature trigger are different for each backdoor target. The attacker changes the labels of the selected samples into corresponding backdoor targets, i.e., y_{t_1} or y_{t_2}. Once the backdoored GNN model is trained with the backdoored training dataset, it is expected that the backdoored model will predict any target node (which can be from an untargeted class) with a feature trigger of value v_1 (v_2) into the target class y_{t_1} (y_{t_2}).

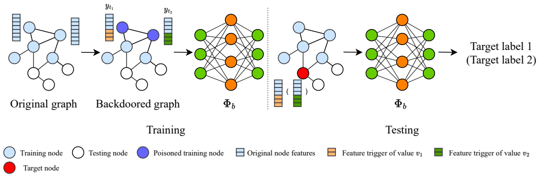

Figure 1: One-to-N backdoor attack framework.

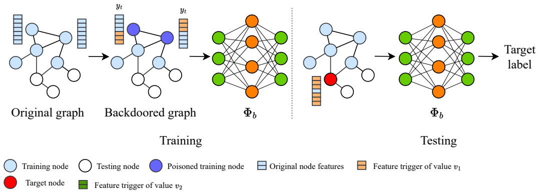

Figure 2: N-to-One backdoor attack framework.

2.2.2 N-to-One Attack. In the N-to-One attack, there are N backdoor triggers $(t_1, t_2, \cdots, t_N)$ and *One* backdoor target (y_t). This attack is only triggered by injecting all the N different backdoor triggers, i.e., only when all the N backdoor triggers are satisfied, the N-to-One attack will misclassify the target node into the target class, while a single backdoor trigger cannot trigger the attack. The attack framework of the N-to-One attack is demonstrated in Figure 2, with $N = 2$ as an example. In the training phase, the attacker samples a subset of the original training dataset, which is not in the target class. For each part of the selected samples, the attacker injects a feature trigger either t_1 or t_2. Unlike the One-to-N attack, in the N-to-One attack, the feature values for all feature triggers are the same, but the injecting position for each trigger is different. For instance, in Figure 2, one feature trigger is injected into the last three dimensions of the feature vector of the node while the other feature trigger is injected into a different place, i.e., the first dimensions of the feature vector. Another difference between the N-to-One and One-to-N attacks is that there is only one backdoor target in the N-to-One attack; thus, the labels of the poisoned samples are all changed to the target label. In the testing phase, the backdoored GNN model is assumed to misclassify any target node with all the backdoor triggers into the target class. It will still output the original label for the target node if only one trigger is present.

3 PRELIMINARY EXPERIMENTAL RESULTS

Our experiments were run on an Intel Core i7-7600U CPU processor with 2.80GHz frequency and 15.5 GiB memory. We used the PyTorch framework and repeated it ten times.

Datasets. We perform experiments with two publicly available real-world datasets for the node classification task: Cora [3] and CiteSeer [3]. These two datasets are citation networks in which each publication is described by a binary-valued word vector indicating the absence/presence of the corresponding word in the collection of 1, 433 and 3, 703 unique words, respectively.

We split 20% of the total nodes as the original training dataset (labeled), and the rest of the nodes are treated as the original testing dataset. To generate the backdoored training dataset, we sample γ of the original training dataset to inject the feature trigger and relabel these nodes with the target label. The trigger size is set to η of the

Table 1: Attack performance (ASR).

Dataset	Model	One-to-One	One-to-N		N-to-One				
			Target 1	Target 2	t_1	t_2	t_3	t_4	t_{all}
Cora	GCN	100.00%	76.61%	99.23%	22.03%	17.87%	15.29%	10.77%	92.25%
	GAT	100.00%	95.78%	99.72%	22.14%	11.29%	11.47%	14.44%	98.15%
CiteSeer	GCN	100.00%	74.66%	98.11%	50.01%	53.26%	47.33%	42.36%	99.60%
	GAT	100.00%	98.30%	96.39%	40.80%	43.41%	34.92%	28.42%	98.51%

total number of node feature dimensions. γ and η significantly impact the backdoor attacks for both One-to-N and N-to-One attacks. We set those parameter values after a tuning phase.

Target Models. We choose GCN [1] and GAT [5] as our target models to be attacked, as these two methods are commonly-used GNN models for the node classification task. We train the clean and backdoored GNN models with a learning rate of 0.005 and use Adam as the optimizer.

Evaluation. We measure the backdoor performance of the model on a fully poisoned dataset $\hat{D}$. The poisoned dataset is generated by embedding the testing dataset of the non-target labels with a trigger, avoiding the influence of the original label. We compute the *attack success rate* $ASR = \frac{\sum_{i=1}^{n} \mathbb{I}(\Phi_b(\hat{x_i})=y_t)}{n}$ where Φ_b is the backdoored model, $\hat{x_i}$ is a poisoned input, i.e., $\hat{x_i} \in \hat{D}$, y_t is the target class, and $\mathbb{I}$ is an indicator function. We use *clean accuracy drop (CAD)* to evaluate the attack evasiveness. CAD indicates the classification accuracy difference between the original GNN model Φ_o and the backdoored GNN model Φ_b over the clean testing dataset.

Results. For the One-to-N attack, we evaluate the attack performance in the case of $N = 2$. In this attack scenario, the feature trigger has two feature values, i.e., $v_1 = 0$, $v_2 = 1$, to launch the attack for two backdoor targets. We randomly select the two backdoor targets between the C classes. In addition, we set the injecting rate γ and trigger size η to be 10% and 5%, respectively. We split the selected training samples into two parts, not uniformly but following the injecting setting of *Weaker_More* [9]. Specifically, the feature trigger with a specific value which is more difficult to be learned by the GNN model is injected with more samples than those with other values. After conducting a tuning phase, we split the selected training samples into two parts in such a way: 2/3 and 1/3 for the trigger of value 0 and 1, respectively, for Cora; 3/4 and 1/4 for the trigger of value 0 and 1, respectively, for CiteSeer.

For the N-to-One attack, we evaluate the attack performance in the case of $N = 4$. Thus, there are four feature triggers of the same value in different locations in the node features. The four injecting locations are selected uniformly at random. It is intuitive that compared to the classical One-to-One attack, the N-to-One attack can achieve a similar attack success rate with a much smaller number of poisoned training samples (less injecting rate) and even a smaller trigger size. Here, we set the injecting rate γ and trigger size η to be 0.5% and 0.5%, respectively. For comparison, we also present the attack performance of the One-to-One attacks.

The backdoor attack results of the One-to-N, N-to-One, and One-to-One attacks are shown in Table 1. For the One-to-N attack, the same backdoor trigger with different values can successfully trigger the corresponding backdoor targets with a high success rate, e.g., 98.30% and 96.39% ASR for targets 1 and 2, respectively, for CiteSeer on the GAT model. The ASR for target 2 is generally higher than that for target 1, which demonstrates that it is easier for the GNN model to learn the trigger of feature value $v_2 = 1$ and further verifies the necessity of the *Weaker_More* injecting setting. For the N-to-One attack, when only a single backdoor trigger is present, the average ASR is lower than 30% in most cases. However, when all the four backdoor triggers ($t_{all} = t_1 \& t_2 \& t_3 \& t_4$) are all satisfied, the ASR can achieve more than 90% for all datasets and models, and it can as high as 99.60% for CiteSeer on GCN model. One-to-N and N-to-One attacks can have a low CAD (around 0.8%), indicating that both attacks have a negligible impact on the original task of the model. Although the ASR of both attacks is lower than that of the One-to-One attack (ASR of 100%), there are significant advantages to these two attacks. In the One-to-N attack, the attacker can trigger multiple backdoor targets. Even if the defender detects one of the N backdoors, he will not realize there are other backdoors with different feature values; the attacker can still trigger the backdoor attacks. For the N-to-One attack, the attacker only requires to inject a much smaller number of training samples than the One-to-One attack, i.e., 0.5% vs. 10%, to achieve a high ASR. Thus, the defenders cannot reverse all the triggers easily.

4 ACKNOWLEDGMENTS

This work is supported by the China Scholarship Council (CSC).

REFERENCES

[1] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[2] Xiang Ling, Lingfei Wu, Wei Deng, Zhenqing Qu, Jiangyu Zhang, Sheng Zhang, Tengfei Ma, Bin Wang, Chunming Wu, and Shouling Ji. 2022. Malgraph: Hierarchical graph neural networks for robust windows malware detection. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1998–2007.

[3] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI magazine* (2008).

[4] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2018. Clean-label backdoor attacks. (2018).

[5] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[6] Zhaohan Xi, Ren Pang, Shouling Ji, and Ting Wang. 2021. Graph backdoor. In *30th USENIX Security Symposium (USENIX Security 21)*. 1523–1540.

[7] Jing Xu, Rui Wang, Stefanos Koffas, Kaitai Liang, and Stjepan Picek. 2022. More is better (mostly): On the backdoor attacks in federated graph neural networks. In *Proceedings of the 38th Annual Computer Security Applications Conference*. 684–698.

[8] Jing Xu, Minhui Xue, and Stjepan Picek. 2021. Explainability-based backdoor attacks against graph neural networks. In *Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*. 31–36.

[9] Mingfu Xue, Can He, Jian Wang, and Weiqiang Liu. 2020. One-to-N & N-to-one: Two advanced backdoor attacks against deep learning models. *IEEE Transactions on Dependable and Secure Computing* 19, 3 (2020), 1562–1578.

[10] Jiawei Zhang, Bowen Dong, and S Yu Philip. 2020. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th international conference on data engineering (ICDE)*. IEEE, 1826–1829.

[11] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. 2021. Backdoor attacks to graph neural networks. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*. 15–26.