Towards Extracting Graph Neural Network Models via Prediction Queries (Student Abstract)

Bang Wu, Shirui Pan, Xingliang Yuan
Monash University, Melbourne, Australia
{bang.wu, shirui.pan, xingliang.yuan}@monash.edu

Abstract

Graph data has been widely used to represent data from various domain, e.g., social networks, recommendation system. With great power, the GNN models, usually as valuable properties of their owners, also become attractive targets of the adversary who covets to steal them. While existing works show that simple deep neural networks can be reproduced by so-called Model Extraction Attacks, how to extract a GNN model has not been explored. In this paper, we exploit the threat of model extraction attacks against GNN models. Unlike ordinary attacks which obtain model information via only the input-output query pairs, we utilize both the node queries and the graph structure to extract the GNNs. Furthermore, we consider the stealthiness of the attack and propose to generate legitimate queries so the extraction can be applied discreetly. We implement our attack by leveraging the responses of these queries, as well as other accessible knowledge, e.g., neighbor connectives of the queried nodes. By evaluating over three real-world datasets, our attack is shown to effectively produce a surrogate model with more than 80% equivalent predictions as the target model.

Introduction

Graph Neural Network models have been widely deployed and achieved high performance in a variance of applications. These well-trained models are highly costly during the data collecting and training and often considered as intellectual property. Consequently, lots of AI platforms, e.g., Amazon SageMaker and Google Cloud AutoML, provide privatization deployments for the model owner to sell their models with the license fee. Such commercialization draws much attention to the security of the models. Previous researches (Tramèr et al. 2016; Oh et al. 2018) have shown that attackers can steal several types of ML models by so-called Model Extraction Attack. They generate a sequence of queries to the target model and reconstruct a similar model derived from the input-output pairs of these queries. However, most of existing attacks only target at the models with non-graph structures and thus the threat to GNN models is still unclear.

Comparing with attacks to other neural networks, model extraction attacks against GNNs introduce new challenges.

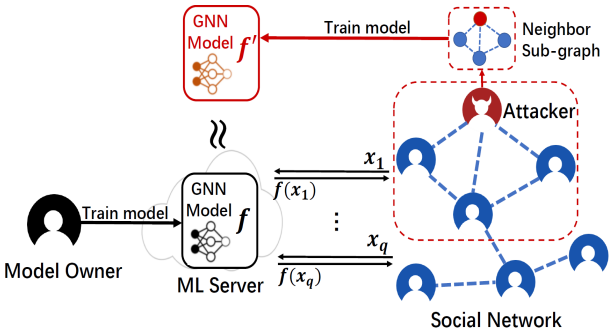

Figure 1: Model Extraction Attacks on a social network. A model owner provides a GNN model f and the service of prediction queries. An attacker obtains its neighbor subgraph to extract a surrogate model $f' \approx f$.

Firstly, unlike the models with non-graph structures, training a GNN model requires not only the input-output pair but also the graph structure. While the target model is trained based on the whole graph, the queries for the duplicated model training cannot cover every node of the target. So, whether and how to extract the model for an entire graph by the queries only from its partial nodes is yet to be explored. Secondly, while the ordinary extractions elaborate a sequence of input queries, it is naturally to also generate a set of node attributes and different graph structures for the attack to the GNN models. However, since these attributes and connectives of the nodes in a graph are highly related, incompatible queries of them can be easily detected in real-world applications. Therefore, the stealthiness of the queries should be considered for the attacks against GNNs.

In this paper, we introduce a model extraction attack against GNNs which can reconstruct a model with similar functionality. The attacker can extract the models by generating the query that looks legitimate, and obtaining accessible knowledge from other nodes in the target graph. Figure 1 shows an example of how our attack steals a GNN model on a social network. We consider an attacker blending in with the normal users in the network. Rather than querying the target model with irregular attributes or modifying the graph structure, the attacker performs like a normal user and only gathers information from his neighbours. The attacker's goal

is to extract a similar GNN model as the target based on their queries and neighbour connectives.

Proposed Attack

Problem Definition: Considering a GNN model $f(.)$, our attack aims to reconstruct a surrogate model $f'(.)$ such that $\forall v_i \in \mathcal{V}, f'(v_i) = f(v_i)$ or $f'(v_i) = y_i$, where $\mathcal{V}$ is a set for all the node in the graph, (v_i, y_i) is the input and correct label pair of a node in $\mathcal{V}$.

Attack Assumptions: Our attack targets at node classification models which have been most commonly used and studied. We also consider a prevalent scenario when the target GNN models only provide black-box access, so they only return the final prediction labels rather than the confidence values, i.e., probability of the output labels. We presume a set of attack nodes controlled by the attacker mingled among the normal nodes in a graph who can query the target GNN model. And we further assume that the target model is alert to the anomalous attribute queries and connective modifications, which can be detected by comparing the graph structure or the feature co-occurrence graph (Zügner, Akbarnejad, and Günnemann 2018). In other words, the attacker can only generate legitimate queries as the normal nodes. Besides obtaining their query responses, the attacker can also obtain the neighbour connectives of the controlled nodes which are often public or accessible in practice.

Attack Method: To train the surrogate model, our approach is to produce a surrogate training dataset based on the attacker's knowledge and the obtained query responses. The attacker randomly selects and controls nodes (called attack nodes) among the graph so the information from the entire graph can be gathered. However, the selected attack nodes might not be connected with each other so they cannot reflect the correlation among the target graph. To deal with this isolation, we propose to synthesize nodes to bridge them and generate structural training graph. Their connectives can be directly referring to the accessible neighbour connections, and their attributes can also be synthesized based on the known attack nodes. The following equation shows how we generate the attributes of the synthetic nodes by considering their 2-hop neighbours:

$$\mathcal{F}_t = \sum_{i=1}^{n} \mathcal{F}_{1-hop\_i}/(n \cdot d_i) + \alpha \cdot \sum_{i=1}^{m} \mathcal{F}_{2-hop\_i}/(m \cdot d_i) \quad (1)$$

where $\mathcal{F}_{k-hop\_i}$ represents the ith k-hop (we select $k = 1$ or $k = 2$) node of the targets, d_i represents the degree of this neighbour, and α is a factor to adjust the effects from 1 or 2 hops nodes. Based on the Six Degrees of Separation theory, most of the isolated attack nodes can be connected via these synthesized nodes. As a result, the complement structural graph, the attributes for every nodes, and a set of attack nodes labeled by their query responses can be generated. And they can be fed into a GNN model in the training phase to derive the duplicated model.

Experiments on Node Classification

We deploy our attack on three public datasets, including Cora, Citeseer and Pubmed. The target GNN models are chosen to be the two-layer graph convolutional network (Kipf and Welling 2017) which is a common GNN architecture. Our attack is assessed by fidelity (Juuti et al. 2019), i.e., the percentage of predictions from the surrogate model the same as the targets.

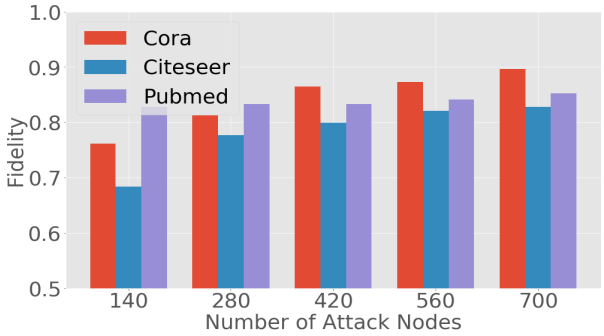

Figure 2: Average fidelity for the attacks with different number of attack nodes.

We evaluate the effectiveness of the attack under various numbers of the attack nodes as shown in Figure 2. The results show that our attack achieves about 90% fidelity with 700 queries in Cora and more than 80% in other two datasets. It can also be found that controlling more attack nodes can increase the attack performance at the beginning. But the improvement will enter the saturation when the number becomes larger.

Conclusions

In this paper, we demonstrate a model extraction attack against GNNs. We first generate legitimate-looking queries as the normal nodes among the target graph, then utilize the query responses and accessible structure knowledge to reconstruct the model. Our experimental results show that our attack obtains surrogate models with similar predictions as the targets. We hope our preliminary work can elicit much attention on the privacy of GNN models. Our future work is to adaptively design the attacks that consider the adversaries with different knowledge to the target graph.

References

Juuti, M.; Szyller, S.; Marchal, S.; and Asokan, N. 2019. PRADA: Protecting Against DNN Model Stealing Attacks. In *EuroS&P 2019*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR 2017*.

Oh, S. J.; Augustin, M.; Fritz, M.; and Schiele, B. 2018. Towards Reverse-Engineering Black-Box Neural Networks. In *ICLR 2018*.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security 16*.

Zügner, D.; Akbarnejad, A.; and Günnemann, S. 2018. Adversarial Attacks on Neural Networks for Graph Data. In *KDD 2018*.