Evaluating the Effectiveness of Defensive Mechanisms Against Model Extraction Attacks in Graph Neural Networks

Gabriella Muñoz, Yushun Dong

FSU UNDERGRADUATE RESEARCH OPPORTUNITY PROGRAM
CENTER FOR UNDERGRADUATE RESEARCH & ACADEMIC ENGAGEMENT

FSU | FLORIDA STATE UNIVERSITY

Introduction

Machine learning models are widely used in real-world and proprietary applications, making them valuable targets for attackers. One key threat is model extraction, where an attacker copies a trained model by querying it and observing its outputs. In graph neural networks (GNNs), which analyze graph-structured data such as social networks or molecular graphs, explanation tools designed to improve transparency may unintentionally reveal sensitive model information.

This project examines explanation-guided model extraction in GNNs and explores whether defensive strategies can reduce information leakage while preserving model performance.

Abstract

- Model extraction attacks threaten machine learning systems by allowing attackers to copy deployed models through limited queries. In graph neural networks (GNNs)—models that learn from network-structured data such as social or molecular graphs—explainability tools may unintentionally reveal sensitive model information.
- This study reproduces an explanation-guided extraction framework that aligns surrogate model training with target model explanations.
- The attack was implemented using PyTorch and PyTorch Geometric and evaluated on graph classification datasets.
- Results show that incorporating explanation alignment significantly improves extraction success compared to standard query-based methods.
- These findings establish a strong baseline and highlight the need for defenses that reduce information leakage while preserving model accuracy.

Methodology

Experimental Setup

- This study reproduced an explanation-guided model extraction attack on graph neural networks using the same datasets and experimental conditions reported in prior work.

Procedure

1. A target model was first trained to perform a graph classification task.
2. The trained model was then queried to collect its predictions and explanation information.
3. This information was used to train a second model (surrogate) designed to imitate the target model's behavior.

Evaluation

- The surrogate model was compared to the target model to measure how closely it replicated the original predictions. Model accuracy and prediction similarity were used to evaluate extraction success and verify whether results aligned with the original study.

Results

The reproduced experiments closely matched the results reported in the original study. Surrogate models trained with explanation information more accurately replicated target model predictions than baseline extraction methods. Overall, extraction success significantly improved when explanation data was incorporated.

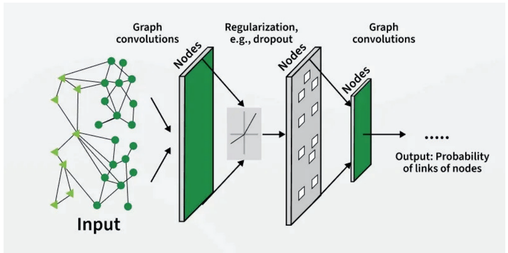

Figure 1: Diagram of a GNN

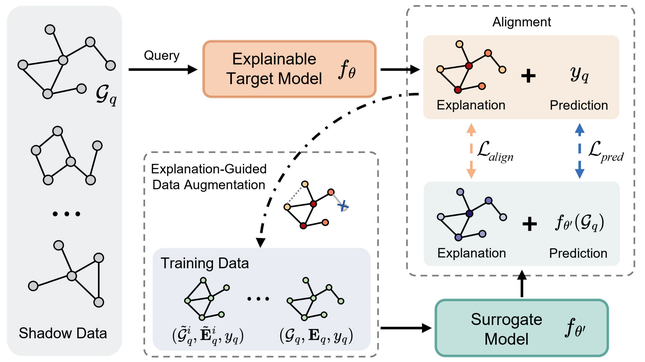

Figure 2: EGSteal Framework

Dataset	Target Acc	Target AUC	Surrogate Acc	Surrogate AUC	Fidelity	Order Acc	Rank Corr
NCI109	0.7321	0.8013	0.6836	0.7529	0.7406	0.6089	0.1983
AIDS	0.9075	0.9574	0.8400	0.9073	0.8775	0.7031	0.3850
Mutagenicity	0.7889	0.8409	0.7589	0.8075	0.8131	0.6912	0.3472

Figure 3: Master table comparing the reproduced results against the original target and surrogate performance for all evaluated datasets.

Discussion

- Reproduced results confirm prior research: explanation information significantly improves model extraction effectiveness in GNNs.
- Interpretability tools, while designed for transparency, may unintentionally expose sensitive model behavior.
- Strength: Successful reproduction of previously reported results under consistent experimental conditions.
- Limitation: Defensive mechanisms were not implemented or evaluated in this study.
- These findings emphasize the need for protective strategies that reduce information leakage while preserving model performance.

References

Scan QR code to view references.

Acknowledgements

I would like to thank my research mentor Yushun Dong, as well as Bolin Shen for their guidance and support throughout this project.