Asim Waheed, Vasisht Duddu, N. Asokan

UNIVERSITY OF WATERLOO

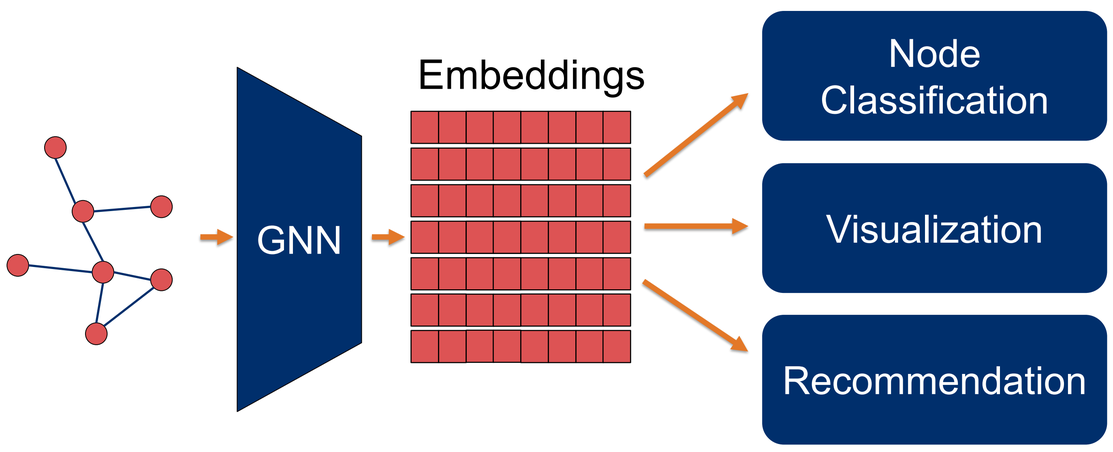
Aalto University

GrOVe: Ownership Verification of Graph Neural Networks using Embeddings

Motivation

- Graph Neural Networks (GNNs) are state-of-the-art on graph data
- Model extraction of GNNs is a realistic threat[1]

How can we design an ownership verification technique for GNNs?



Desiderata

Effective: Separates surrogate and independent

Robust: Resists attempts to circumvent

Efficient: Reasonable computational overhead

Non-Invasive: No utility drop for target model

Intuition

- Embeddings are unique for each input graph
- Surrogate and target embeddings are similar

Can GNN embeddings be used as fingerprints?

Adversary Model

Blackbox access to embeddings[2]

- Type 1: Knows graph structure and features
- Type 2: Estimates adjacency matrix

Goal: Train surrogate with high

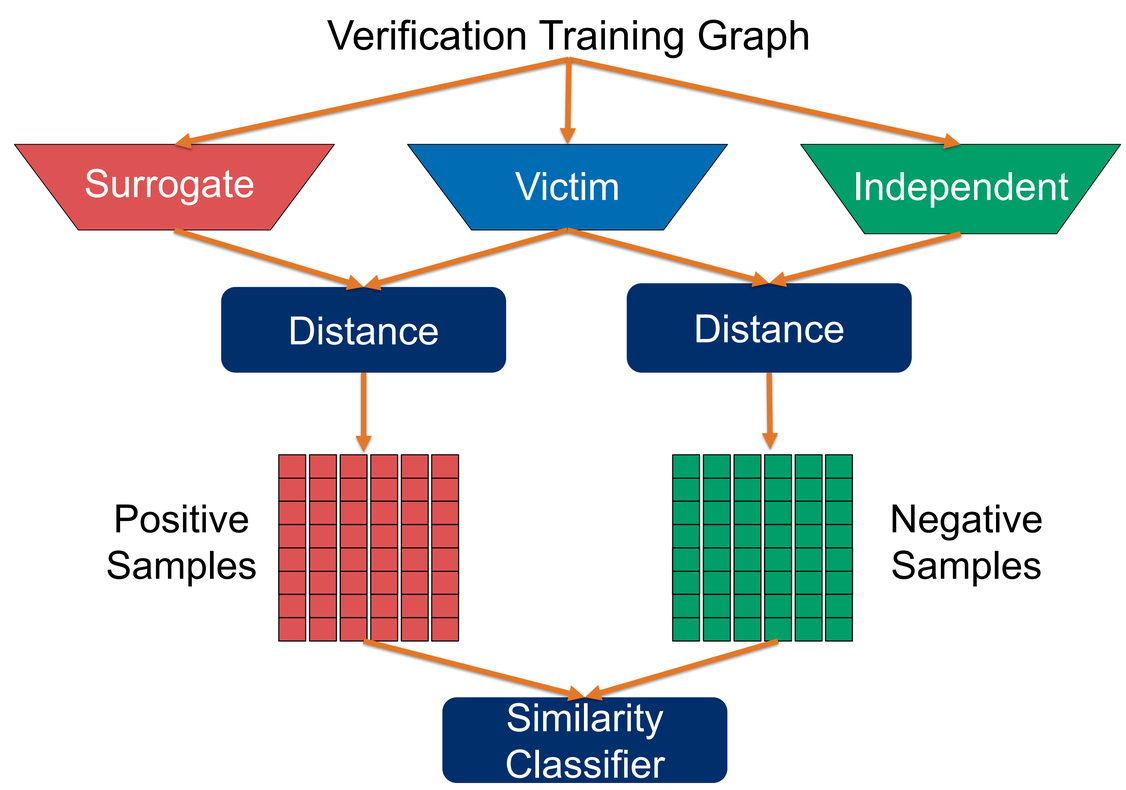
- accuracy on primary task
- fidelity with target
- surrogate closer to independent than target

Verifier

- Sample verification graph data from target's data distribution (avoids false claims[2])
- Blackbox access to target and suspect

Approach

1. Use verification graph to generate embeddings from target and suspect
2. Train a similarity classifier to classify whether embeddings are distinct or similar



3. Verification: Use similarity classifier on suspect and target embeddings to decide

Results

- Zero false-positives / false-negatives across different models for both attacks (effective)
- Robust against fine-tuning, double-extraction
  → Adversarial training to mitigate pruning
- Reasonable cost for verifier (efficient)
- No modification of target (non-Invasive)

[1] Shen et al. *Model Stealing Attacks against Inductive Graph Neural Networks*. IEEE SP 2022.

[2] Liu et al. *False Claims against Model Ownership Resolution*. USENIX Sec 2024.

ssg-research.github.io/mlsec/modelExtDef