Boosting the Adversarial Robustness of Graph Neural Networks: An OOD Perspective

Kuan Li, Yiwen Chen, Yang Liu, Jin Wang, Qing He, Minhao Cheng, Xiang Ao

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
PennState
UCLA Samueli School of Engineering

Introduction

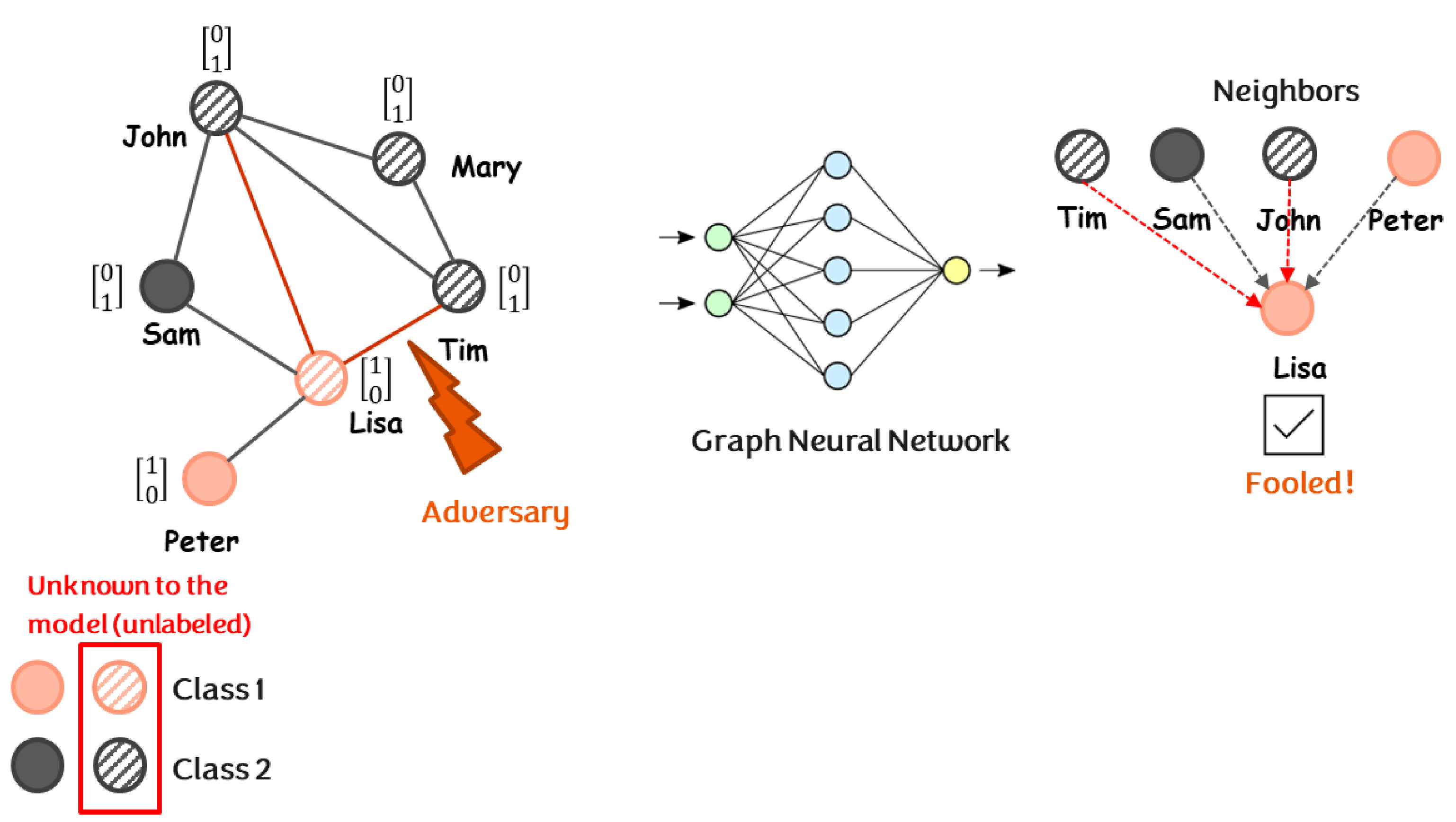
TL;DR: Robust graph neural networks can be easily bypassed by adaptive attacks. How can we achieve adaptive robustness on graphs?

- Based on the evaluation of typical adversarial training, we employ a novel paradigm that leverages the adversarial samples to enhance robustness.
- Through the lens of OOD, we re-examine graph attacks and defenses and, for the first time, propose the existence of a trade-off between the effectiveness and defensibility of attacks in the context of graph adversarial attacks.
- We conduct extensive experiments to compare our methods with other baselines in adaptive and non-adaptive settings.

Graph Adversarial Attacks

The attacker's objective is to find an optimal perturbed graph $\hat{\mathcal{G}}$ that maximally impairs the overall performance of the downstream classifier. This can be formulated as follows

$$\underset{\hat{\mathbf{A}} \in \Phi(\mathbf{A})}{\operatorname{argmin}} \quad \mathcal{L}_{atk}(f_{\theta^*}(\hat{\mathbf{A}}, \mathbf{X}), \boldsymbol{y}),$$



Traditional Adversarial Training

Previous robust GNNs rely on specific properties, so the adversary can easily defeat the defenses by imposing constraints on the same properties during the attack.

What about adversarial training?

$$\mathrm{R}_{\mathrm{ADV}}(\theta) = \left[\max_{x' \in \mathcal{B}(x)} \mathbb{E}_{p_d(x,y)} \mathcal{L}_{CE}(x', y; f_\theta)\right] = \mathbb{E}_{p_d(x)} \max_{x' \in \mathcal{B}(x)} - \left[\sum_y p_d(y \mid x) \log p_\theta(y \mid x')\right]$$

Structural adversarial training

$$\mathrm{R}^{\mathcal{G}}_{\mathrm{ADV}}(\theta) = \max_{\hat{\mathbf{A}} \in \mathcal{B}(\mathbf{A})} \mathbb{E}_{p_d(\mathcal{G}=\{\mathbf{A},\mathbf{X}\})} \mathcal{L}_{CE}\left(f_\theta\left(\hat{\mathbf{A}}, \mathbf{X}\right), \boldsymbol{y}\right)$$

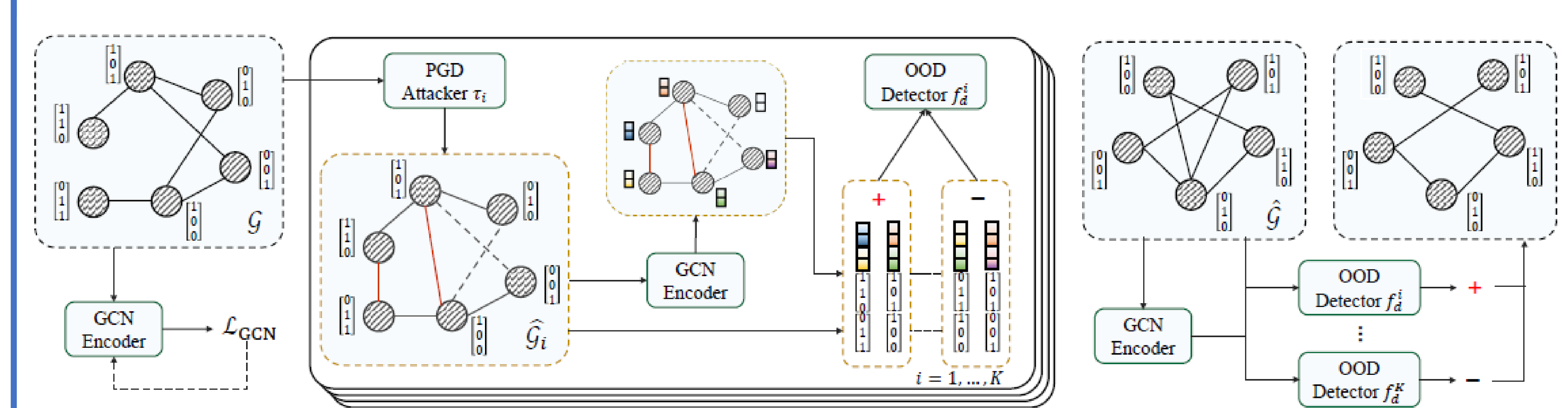
$$= \max_{\hat{\mathbf{A}} \in \mathcal{B}(\mathbf{A})} -\mathbb{E}_{p_d(\mathcal{G})}\left[\sum_y p_d(y \mid x, \mathcal{S}_x) \log p_\theta(y \mid x, \hat{\mathcal{S}}_x)\right]$$

The model will learn incorrect mapping relationships.

Our Solution: GOOD-AT

- Perturbations on images are **continuous and indistinguishable**.
- Perturbations on graphs are **discrete and separated** from clean edges so that they are **removable.**

We train an ensemble OOD detector to remove adversarial edges.



Graph **OOD** Detection-based **A**dversarial **T**raining
GOOD-AT

Adaptive Robustness

Two adaptive attacks against GOOD-AT

- **Resample -** if the sampled adversarial edge generated by PGD can be detected by the detectors, it is discarded
- **Regularization**

$$\mathcal{L}_{all} = \mathcal{L}_{atk} + \lambda \mathcal{L}_{reg}, \text{where } \mathcal{L}_{reg} = \frac{1}{N^2 K} \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{\mathbf{A}}_{ij} f_d^k(e_{ij})$$

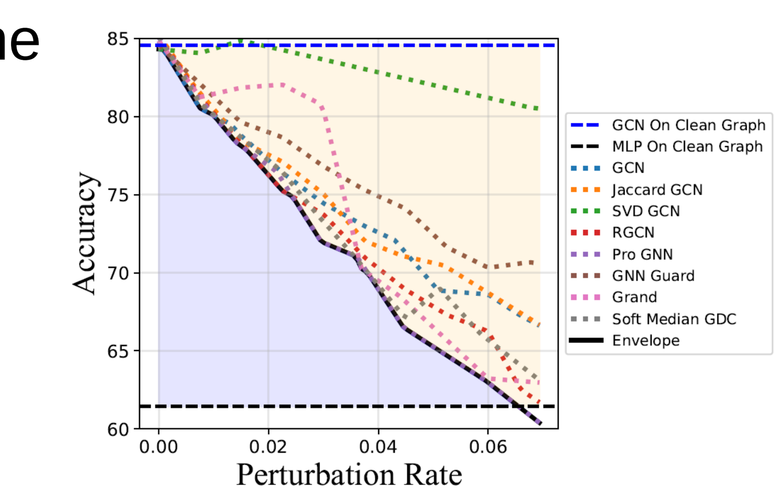
Ptb Rate	3%	6%	9%	12%	15%
PGD-GCN	78.26 ± 1.56	75.10 ± 0.71	72.15 ± 1.45	67.83 ± 1.48	66.39 ± 1.28
PGD-GOOD	84.25 ± 1.90	83.60 ± 1.77	82.71 ± 1.14	82.21 ± 1.73	81.61 ± 1.10
$\mathbf{PGD_{res}}$-GOOD	82.59 ± 1.53	81.06 ± 1.06	79.38 ± 1.15	77.46 ± 1.68	76.54 ± 1.62
$\mathbf{PGD_{reg}}$-GOOD	82.94 ± 1.50	81.72 ± 1.33	80.54 ± 1.66	79.38 ± 2.01	78.63 ± 1.40

Trade-off Between Effectiveness And Defensibility

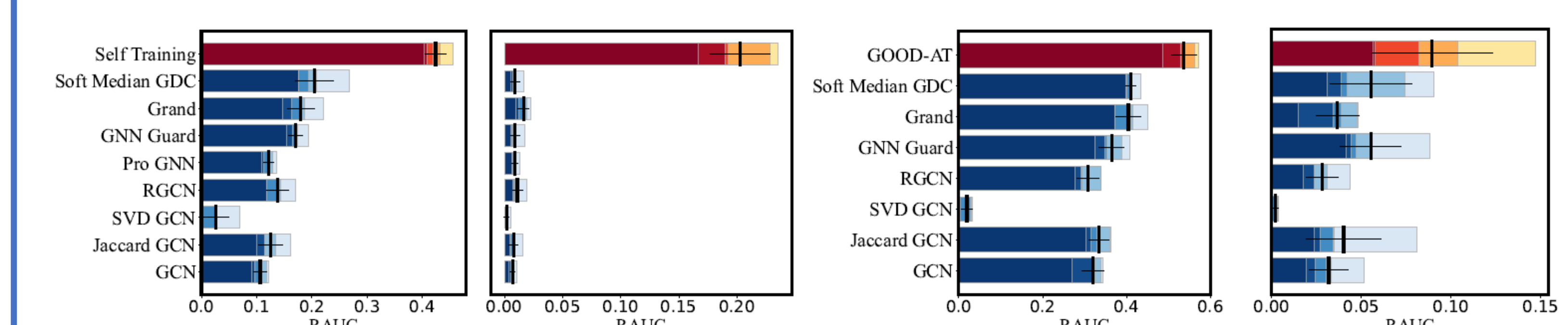
$\mathrm{PGD_{res}}$ degrades to a vanilla GCN, so perturbations that can circumvent detectors are more likely to be in-distribution, which are not that harmful to GNNs.

Adversarial Unit Test

Metric: Relative Area Under the Envelope Curve (RAUC), a budget agnostic metric



$$\mathrm{RAUC}(c) = \int_0^{b_0} (c(b) - a_{\mathrm{MLP}})\, db \text{ s.t. } b \lesssim b_0 \implies c(b) \gtrsim a_{\mathrm{MLP}}$$



(a) Cora, Poisoning (b) Citeseer, Poisoning (c) Cora, Evasion (d) Citeseer, Evasion

ICLR