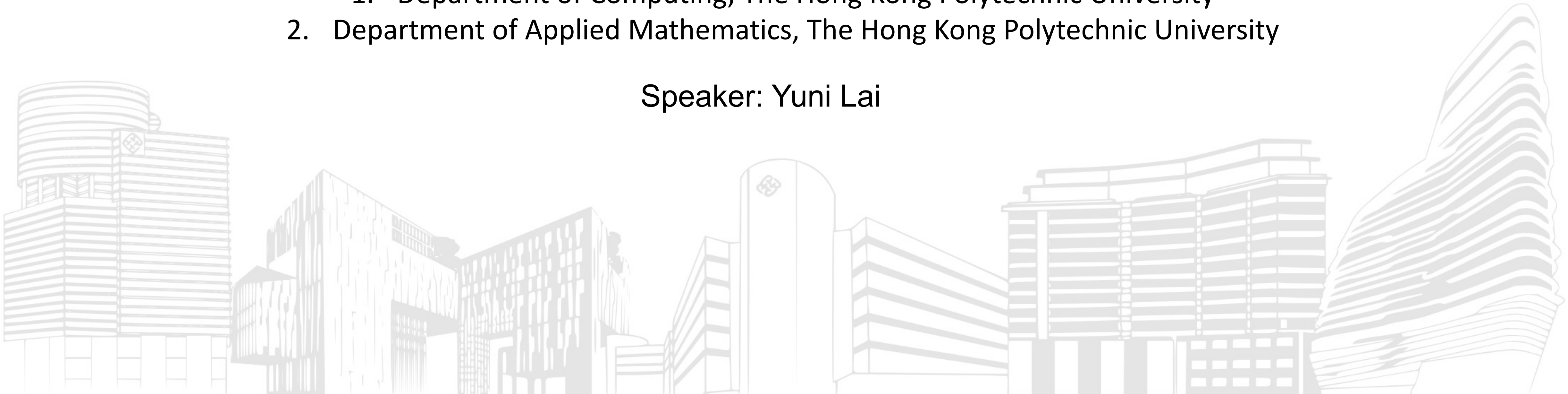
Collective Certified Robustness Against Graph Injection Attacks

Yuni Lai [1] Bailin Pan [2] Kaihuang Chen [2] Yancheng Yuan [2] Kai Zhou [1]

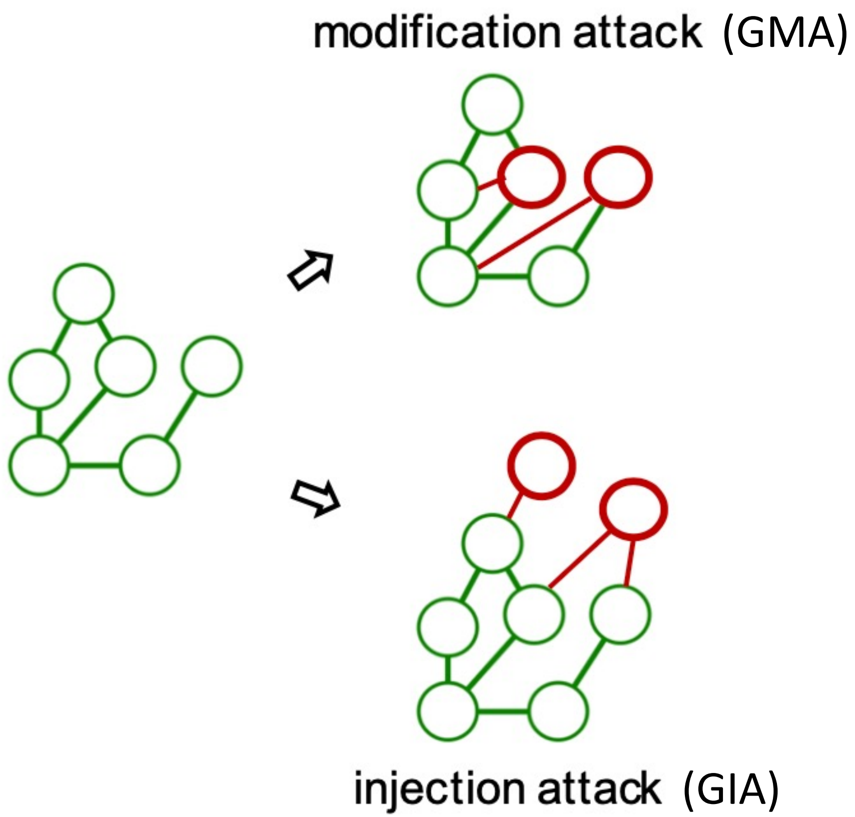
1. Department of Computing, The Hong Kong Polytechnic University
2. Department of Applied Mathematics, The Hong Kong Polytechnic University

Speaker: Yuni Lai



Introduction

Graph Injection Attacks (GIA)



(a) GMA vs. GIA

Zou, Xu, et al. (SIGKDD 2021).

The attacker is able to delete and insert edges among the existing nodes.

Perturb node classification

The attacker injects new nodes, and insert edges from the injected nodes to connect the existing nodes.

For example, in social network, it could be difficult for the attacker to control all the normal users, but it can be easy for the attacker to create a new account, and then interact with normal users.

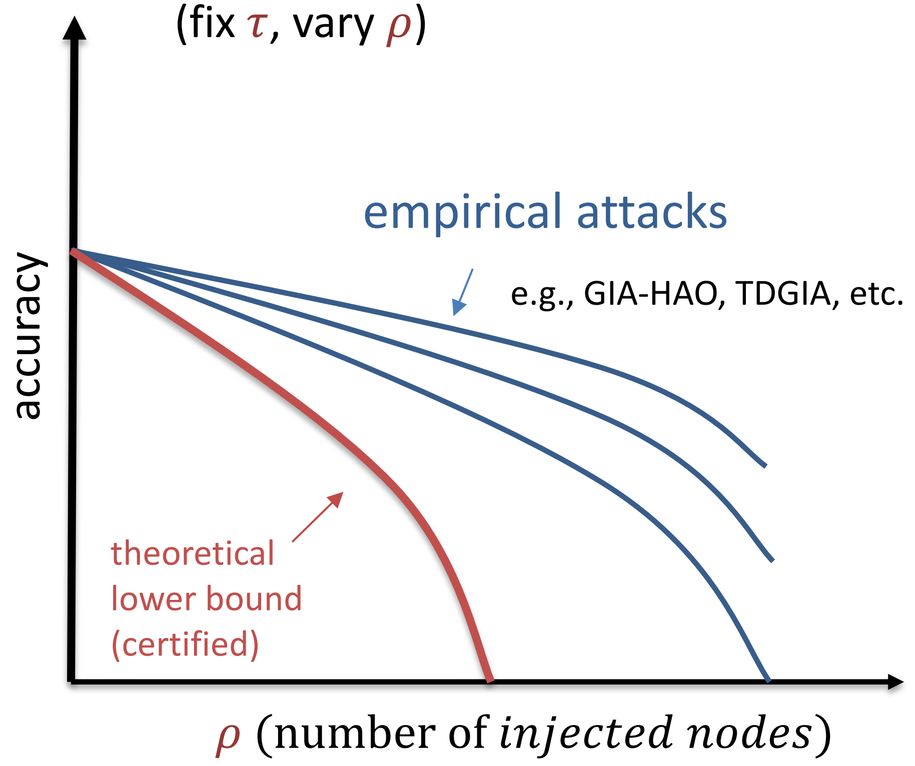
Introduction

Certified Robustness against Graph Injection Attack (GIA)

A node classifier f is certifiably robust for a given input graph G if : we guarantee that the classifier's prediction is consistent within some attack budget $B_{\rho,\tau}(G)$:

$$B_{\rho,\tau}(G) := \{G'(\mathcal{V}',\mathcal{E}',X')|\mathcal{V}' = \mathcal{V} \cup \tilde{\mathcal{V}}, \mathcal{E}' = \mathcal{E} \cup \tilde{\mathcal{E}},$$
$$X' = X \cup \tilde{X}, |\tilde{\mathcal{V}}| \leq \rho, \delta(\tilde{v}) \leq \tau, \forall \tilde{v} \in \tilde{\mathcal{V}}\}$$

The attacker can inject:
ρ malicious nodes, with τ malicious edges per node.



Robustness Certification: lower bound of the model accuracy under a given attack powerρ and τ.

Node-aware Bi-Smoothing (Y. Lai. et al., S&P 2024)

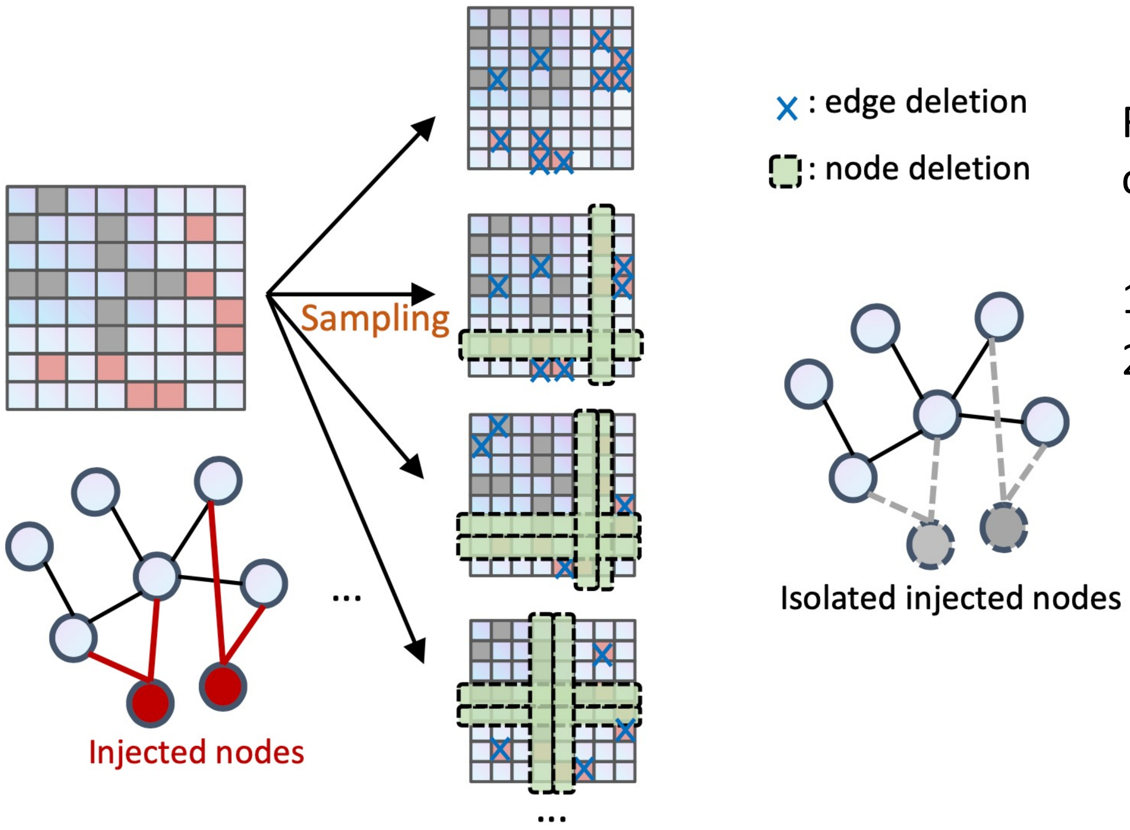


Figure: Node-aware Bi-smoothing (Y. Lai. et al., S&P 2024).

For any node classifier $f(\cdot)$, its smoothed classifier $g(\cdot)$ can be created by:

1. Node and edge randomization: $\phi(G) = \big(\phi_e(G), \phi_n(G)\big)$;
2. Majority vote:

$$g_v(G) \coloneqq \arg max_{y\in\{1,\dots,K\}}\ p_{v,y}(G),$$

$$p_{v,y}(G) \coloneqq P\big(f_v(\phi(G)) = y\big).$$

The goal of certified robustness is to verify :

$$g_v(G) \overset{?}{=} g_v(G'), \forall\ G' \in B_{\rho,\tau}\ (G).$$

Overview

Sample-wise v.s. Collective Certificate

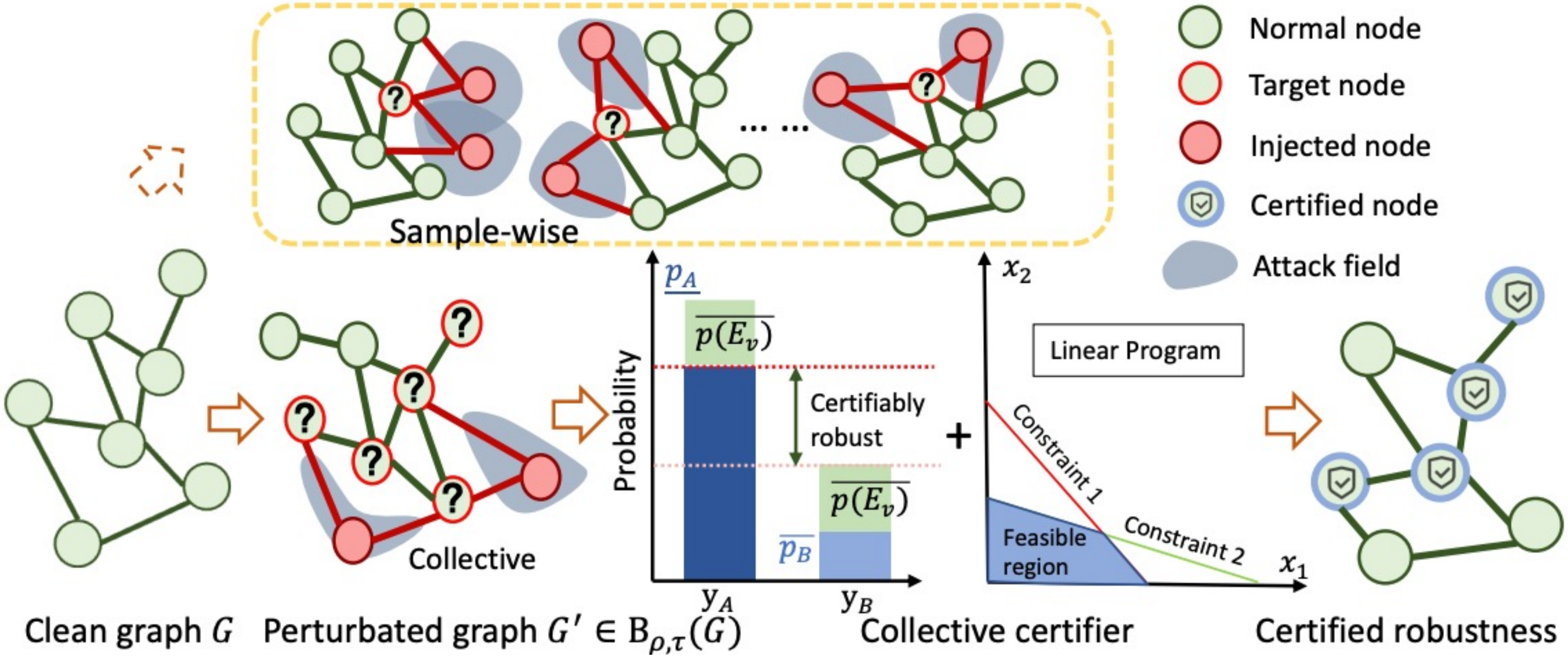
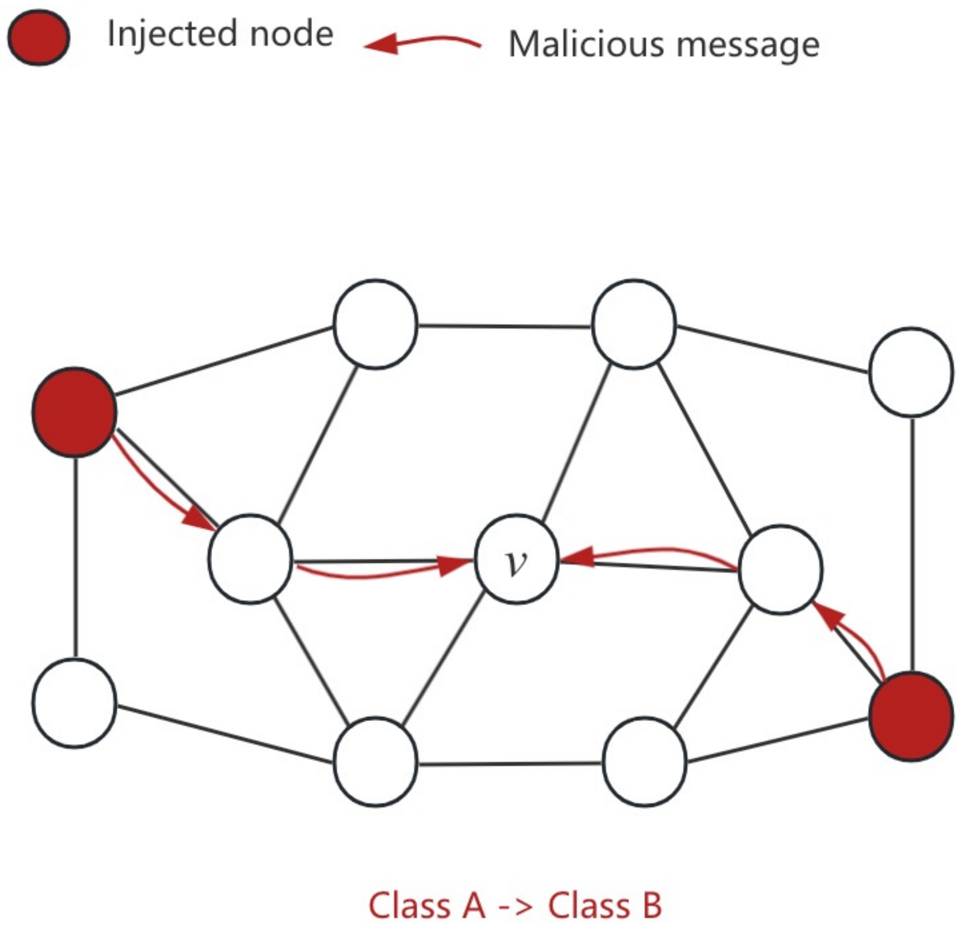


Figure 1: While the sample-wise certificate verifies target nodes one by one, our collective certificate verifies a set of target nodes simultaneously by linear programming.

Our model is inspired by the idea of

Randomized Message-Interception Smoothing (Y. Scholten. et. al. NeurIPS 2022)**.**

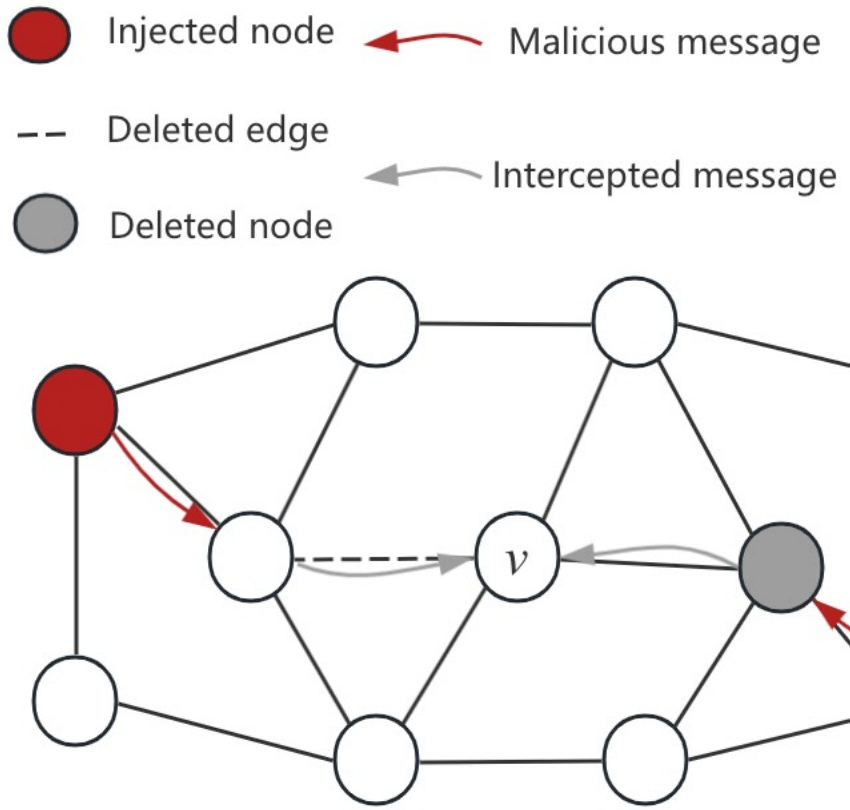
2-layer message passing GNNs



(Figure adapted from Y. Scholten. et. al. NeurIPS 2022)

Malicious message has probability to be intercepted in the Node-aware Bi-smoothing.

2-layer message passing GNNs

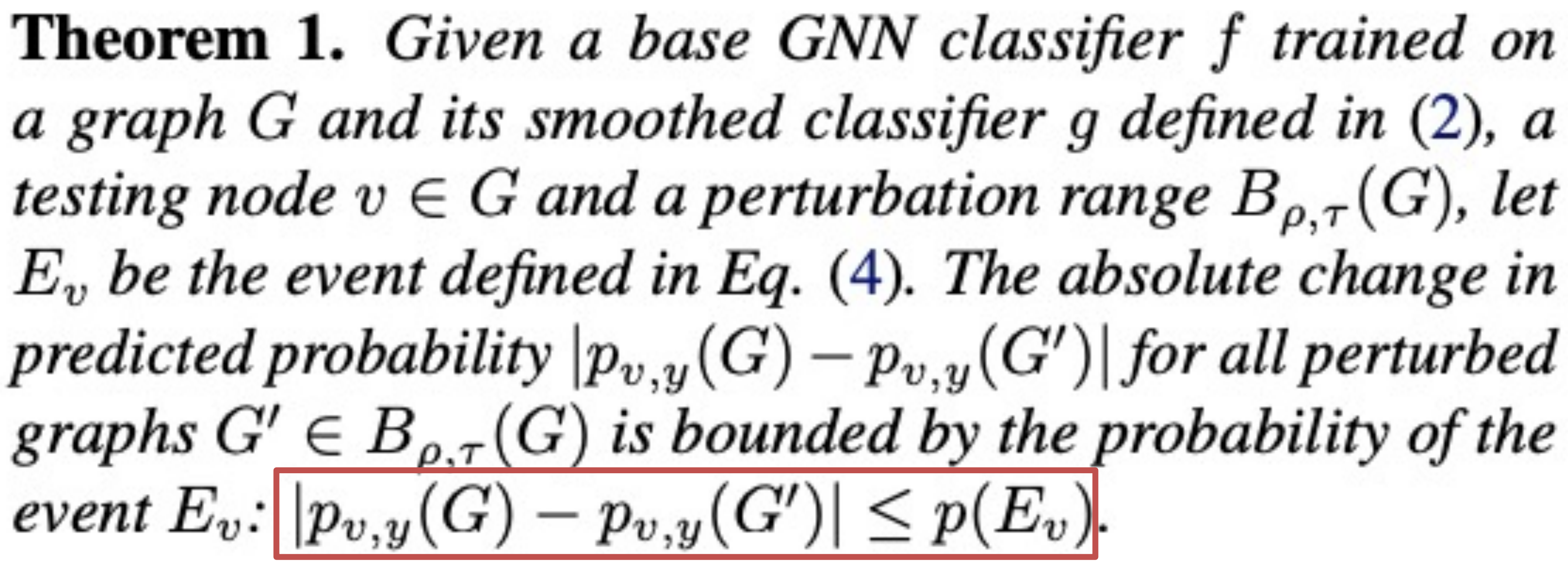


Class A -> Class A

(Figure adapted from Y. Scholten. et. al. NeurIPS 2022)

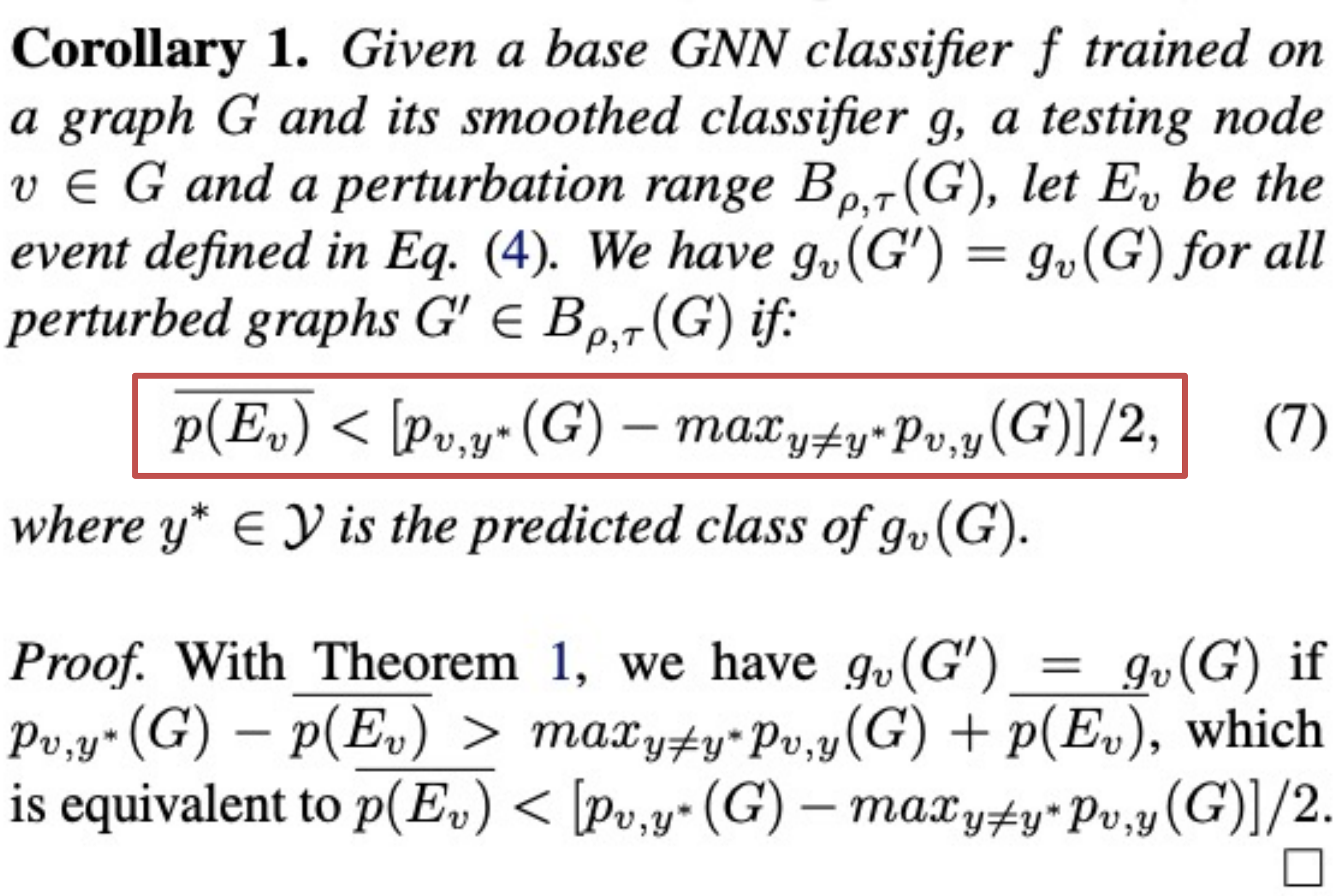
Let $p(E_v)$ denotes the probability that the malicious messages is NOT intercepted, then we have that the absolute change of prediction probability is bounded by $p(E_v)$.

(Because in the other cases, the attacker can not affect the predictions of f .)



Theorem 1. *Given a base GNN classifier f trained on a graph G and its smoothed classifier g defined in (2), a testing node $v \in G$ and a perturbation range $B_{\rho,\tau}(G)$, let E_v be the event defined in Eq. (4). The absolute change in predicted probability $|p_{v,y}(G) - p_{v,y}(G')|$ for all perturbed graphs $G' \in B_{\rho,\tau}(G)$ is bounded by the probability of the event E_v:* $\boxed{|p_{v,y}(G) - p_{v,y}(G')| \leq p(E_v)}$.

Then, we know that the prediction is consistent if:
the upper bound of $p(E_v)$ is smaller than half of the probability gap.



Corollary 1. *Given a base GNN classifier f trained on a graph G and its smoothed classifier g, a testing node $v \in G$ and a perturbation range $B_{\rho,\tau}(G)$, let E_v be the event defined in Eq. (4). We have $g_v(G') = g_v(G)$ for all perturbed graphs $G' \in B_{\rho,\tau}(G)$ if:*

$$\boxed{\overline{p(E_v)} < [p_{v,y^*}(G) - max_{y \neq y^*} p_{v,y}(G)]/2,} \qquad (7)$$

where $y^* \in \mathcal{Y}$ is the predicted class of $g_v(G)$.

Proof. With Theorem 1, we have $g_v(G') = g_v(G)$ if $p_{v,y^*}(G) - \overline{p(E_v)} > max_{y \neq y^*} p_{v,y}(G) + \overline{p(E_v)}$, which is equivalent to $\overline{p(E_v)} < [p_{v,y^*}(G) - max_{y \neq y^*} p_{v,y}(G)]/2$. □

Collective Certificate Original Problem

$B_{\rho,\tau}(G)$: the attacker can inject ρ malicious nodes, with τ malicious edges per node.

$\mathbb{T}$: A set of target nodes.

$$\min_{G' \in B_{\rho,\tau}(G)} \quad |\mathbb{T}| - \sum_{v \in \mathbb{T}} \mathbb{I}\{g_v(G') \neq g_v(G)\},$$

$$\text{s.t.} \quad |\tilde{\mathcal{V}}| \leq \rho,\ \delta(\tilde{v}) \leq \tau,\ \forall \tilde{v} \in \tilde{\mathcal{V}}.$$

(NP-hard)

Upper-bounding the number of non-robust nodes

Collective Certificate Relaxation

$p(E_v)$: the probability that the malicious messages is not intercepted.

$c_v \coloneqq p_{v,y*}(G)$ - $max_{y\neq y*}\, p_{v,y}(G)$.

$$\max_{G' \in B_{\rho,\tau}(G)} \quad M = \sum_{v\in\mathbb{T}} \mathbb{I}\{\overline{p(E_v)} \geq c_v/2\},$$
$$\text{s.t.} \quad |\tilde{\mathcal{V}}| \leq \rho,\ \delta(\tilde{v}) \leq \tau,\ \forall \tilde{v} \in \tilde{\mathcal{V}},$$

The remaining $|\mathbb{T}| - M^*$nodes are certified robust.

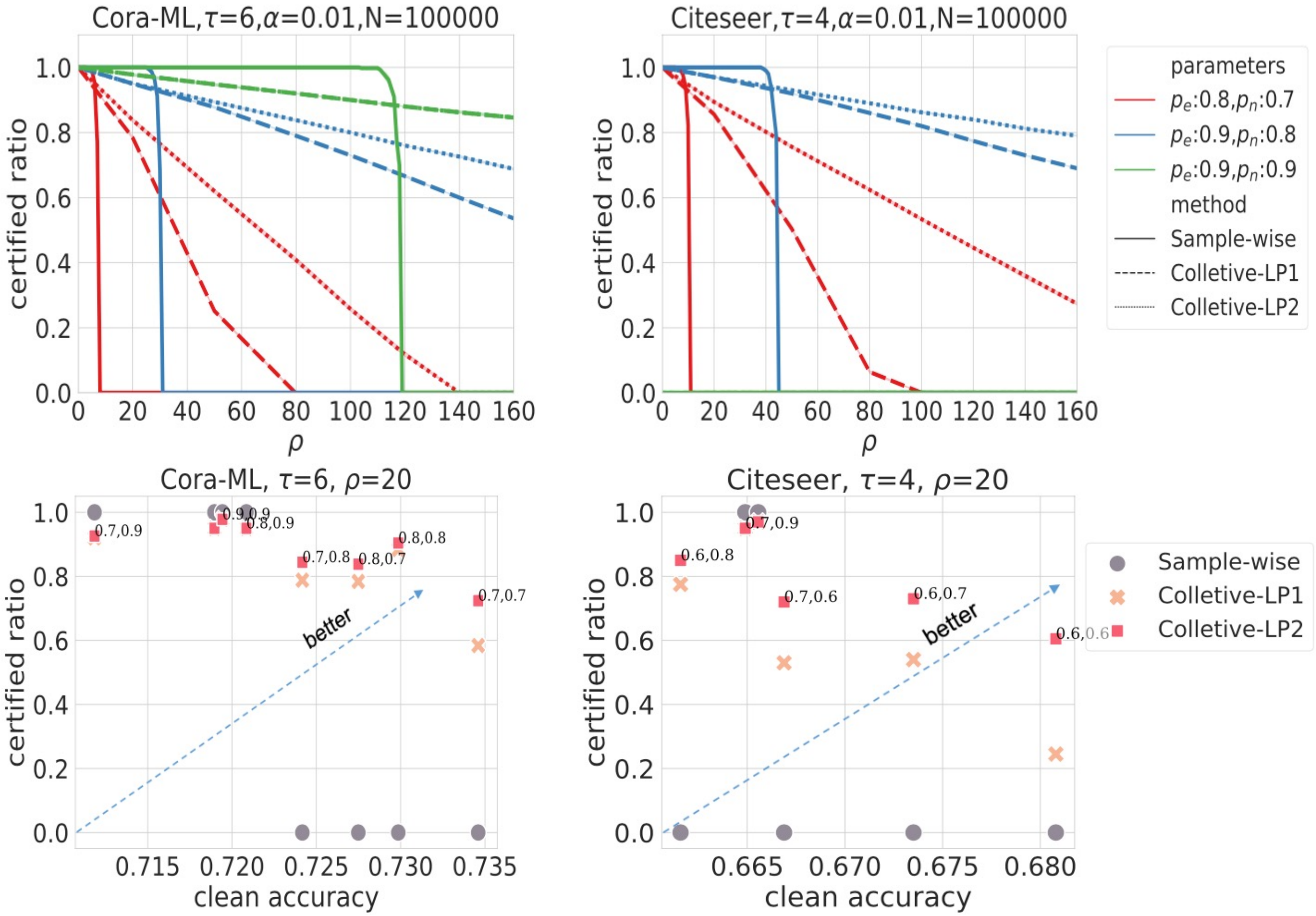
(NP-hard)

To solve the problem, we relax the optimization problem into linear programming.

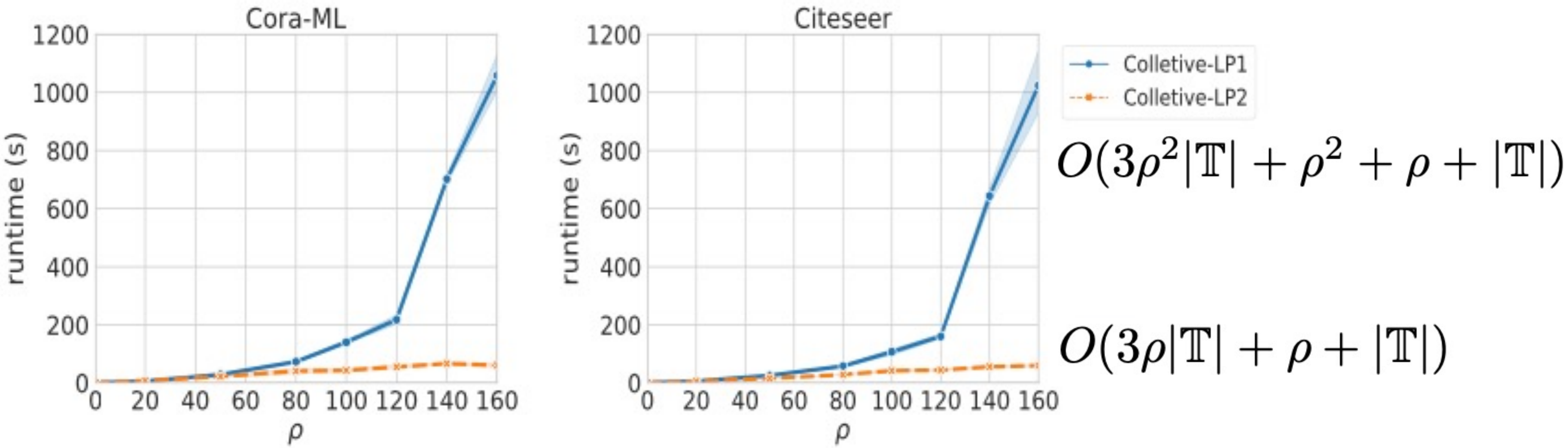
In our paper, we propose two relaxation scheme: Collective-LP1 and Collective-LP2.

Evaluation

Certified ratio and clean accuracy



Evaluation Efficiency



$O(3\rho^2|\mathbb{T}| + \rho^2 + \rho + |\mathbb{T}|)$

$O(3\rho|\mathbb{T}| + \rho + |\mathbb{T}|)$

(a) Runtime

(b) Runtime

Figure 4: Runtime comparison of LP collective models.

Thank you for your attention!

Yuni Lai