



Provably Robust Explainable Graph Neural Networks against Graph Perturbation Attacks

Jiate Li, Meng Pang*, Yun Dong, Jinyuan Jia, Binghui Wang*

*corresponding authors

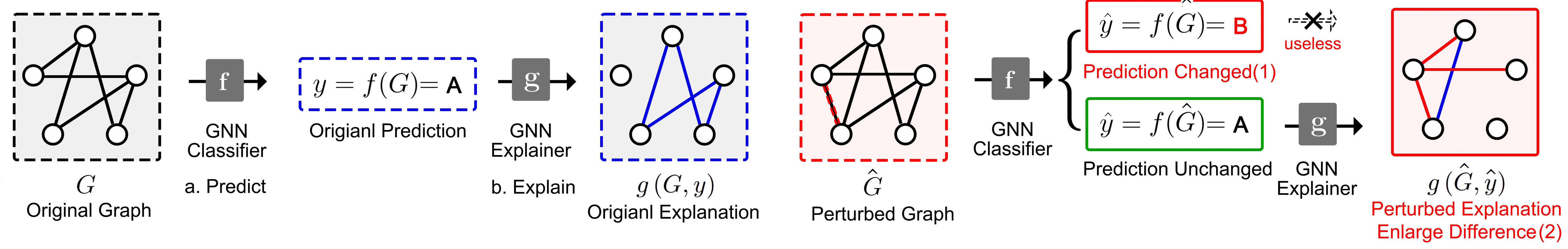


Virtual Page

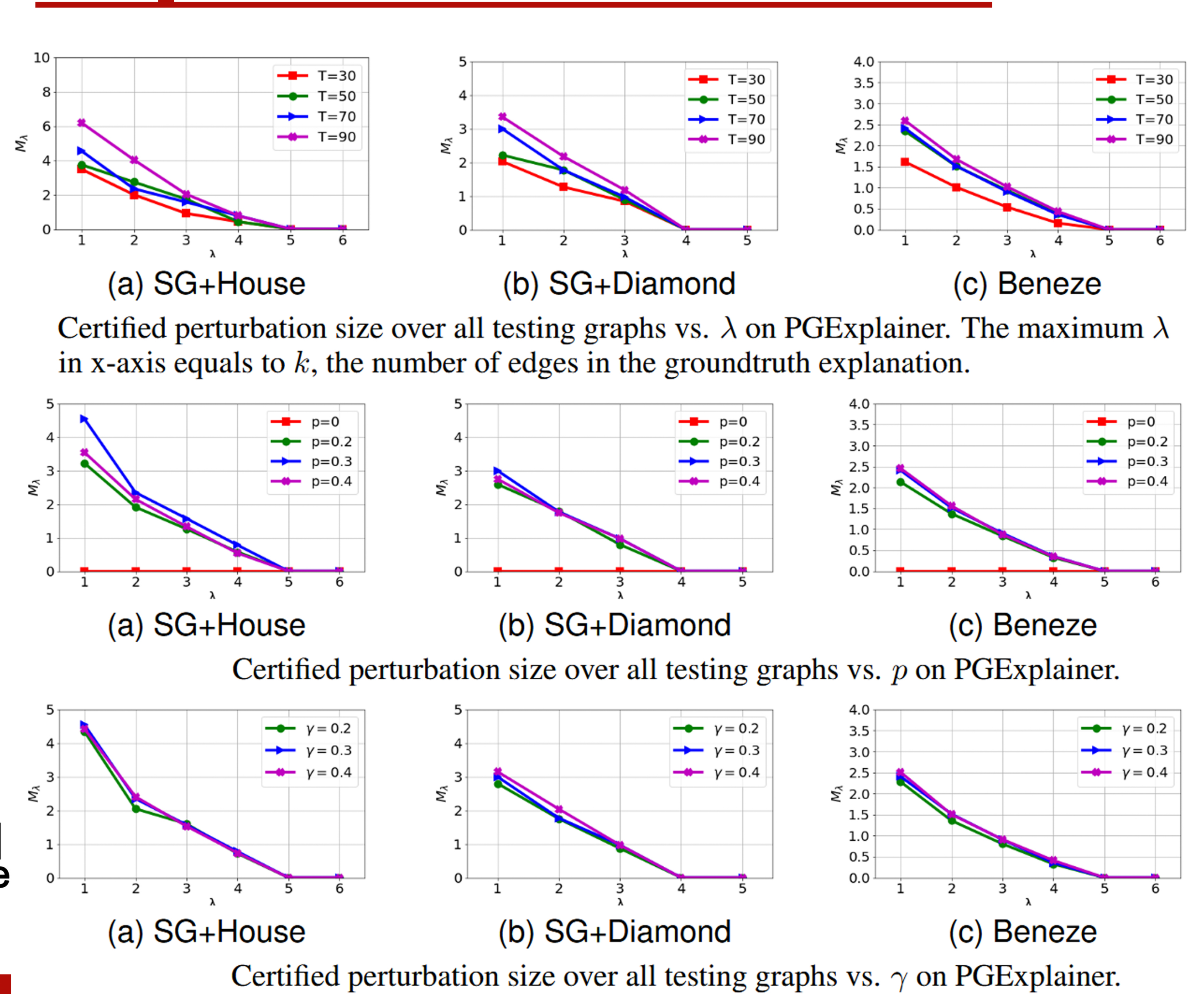


Defending XGNN means building a GNN classifier & explainer system that holds **both** results against adversarial evasion attack.

In this paper, we focus on **edge perturbation attack**. Our method is **provably robust** against **arbitrary** attack with a **deterministic bound**.



Experiments Result



Datasets	Orig.	T					Datasets	Orig.	T				
		30	50	70	90			30	50	70	90		
SG+House	0.740	0.658	0.725	0.725	0.673		SG+House	0.920	0.895	0.905	0.905	0.890	
SG+Diamond	0.745	0.704	0.730	0.729	0.620		SG+Diamond	0.965	0.935	0.935	0.935	0.930	
SG+Wheel	0.629	0.587	0.612	0.571	0.542		SG+Wheel	0.915	0.905	0.905	0.900	0.885	
Benezene	0.552	0.421	0.497	0.468	0.429		Benezene	0.758	0.746	0.700	0.723	0.707	
FC	0.531	0.385	0.452	0.373	0.328		FC	0.711	0.674	0.692	0.692	0.631	

Explanation accuracy on original PGEx and XGNNCert.

Prediction accuracy on original GCN and XGNNCert.

Datasets	Orig.	p				gamma			h		
		0.0	0.2	0.3	0.4	0.2	0.3	0.4	MD5	SHA1	SHA256
SG+House	0.053	0.695	0.695	0.725	0.710	0.715	0.725	0.720	0.725	0.718	0.710
SG+Diamond	0.045	0.620	0.729	0.720	0.712	0.729	0.718	0.729	0.729	0.729	0.721
SG+Wheel	0.042	0.511	0.571	0.508	0.550	0.571	0.564	0.571	0.565	0.565	0.562
Benezene	0.102	0.433	0.468	0.403	0.440	0.468	0.452	0.468	0.472	0.468	0.468
FC	0.096	0.353	0.373	0.288	0.345	0.373	0.385	0.373	0.382	0.390	0.390

Explanation accuracy of our XGNNCert under different p, gamma, and the hash function h

Datasets	SG+House	SG+Diamond	SG+Wheel	Benezene	FC
Exp. Acc.	V-InfoR 0.693	0.419	0.439	0.345	0.217
	XGNNCert 0.740	0.729	0.571	0.468	0.403
Difference Fraction	V-InfoR 48.39%	73.07%	65.35%	83.82%	63.22%
	XGNNCert 7.44%	0.0%	4.20%	1.44%	1.41%

Explanation accuracy and the fraction of different edges under attack in [Li et al.] (2024).

Problem Definition

Adversarial Attack on XGNN

For an XGNN framework with classifier f and explainer g :

$$y = f(G) \quad \mathcal{E}_k = \mathcal{E}.top_k(\mathbf{m}) = \mathcal{E}.top_k(g(G, y))$$

The attacker apply **edge perturbation** on test graph G :

$$\hat{G} = (V, \hat{\mathcal{E}}) \quad \hat{y} = f(\hat{G}) \quad \hat{\mathcal{E}}_k = \hat{\mathcal{E}}.top_k(g(\hat{G}, \hat{y}))$$

Aiming to (1) **alter classification**; or (2) **differ explanation**.

$$\hat{y} \neq y \quad \max |\hat{\mathcal{E}}_k - \mathcal{E}_k|$$

An XGNN is (M_λ, λ) -Certifiably Robust if:

$$\forall \hat{\mathcal{E}} : |\mathcal{E} - \hat{\mathcal{E}}| \leq M_\lambda \rightarrow (y = \hat{y}) \wedge (|\mathcal{E}_k \cap \hat{\mathcal{E}}_k| \geq \lambda)$$

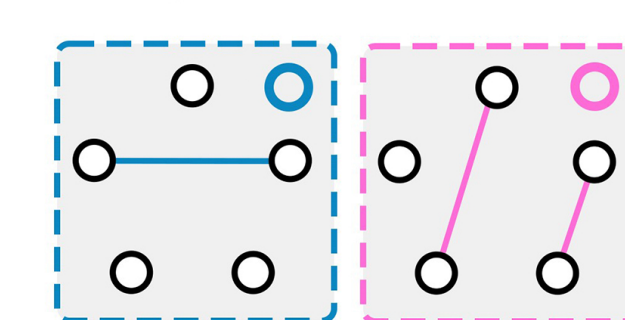
Proposed Method: XGNNCert

Generating Hybrid Subgraphs

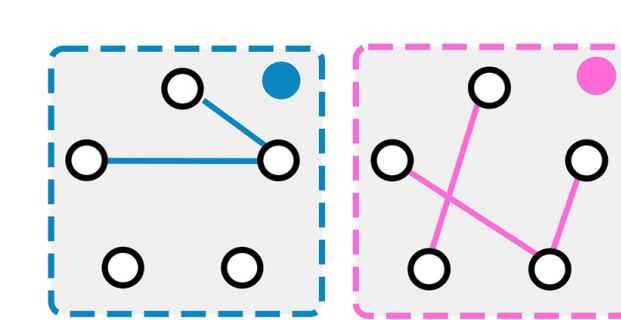
$$i_e = h[\text{str}(u) + \text{str}(v)] \bmod T + 1.$$

Edge Index: hash edges into T groups.

$$\mathcal{E}^i = \{\forall e \in \mathcal{E} : i_e = i\} \quad \mathcal{E}_C^i = \{\forall e \in \mathcal{E}_C : i_e = i\}$$

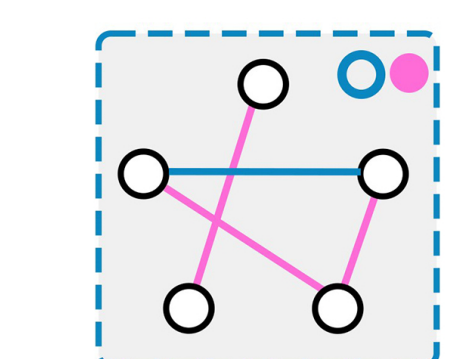


Subgraphs



Complete Subgraphs

$$\mathcal{T}_{-i} = \mathcal{T} \setminus i \xrightarrow[p \text{ fraction}]{\text{random sample}} \mathcal{T}_{-i}^p,$$



$$\mathcal{E}_H^i = (\cup_{j \in \mathcal{T}_{-i}^p} \mathcal{E}_C^j) \cup \mathcal{E}^i$$

Hybrid Subgraphs

Majority Classification & Explanation

Voting-based Classifier: collect **predictions on subgraphs as votes** and predict the **majority class**:

$$n_c = \sum_{i=1}^T \mathbb{I}(f(G_H^i) = c), \forall c \in \mathcal{C}, \quad \bar{f}(G) = \arg \max_{c \in \mathcal{C}} n_c.$$

Voting-based Explainer: collect **top r% explanatory edges in majority-classified subgraphs** as votes:

$$n_e^\gamma = \sum_{i=1}^T \mathbb{I}(m_e^i \geq m_{(\gamma)}^i) \cdot \mathbb{I}(f(G_H^i) = \bar{f}(G)), \forall e \in \mathcal{E}_C,$$

Take **existent edges with top k votes** as explanation:

$$\bar{g}^\gamma(G, \bar{f}(G)) = \mathcal{E}.top_k(\mathbf{n}^\gamma).$$

Certify Deterministic Bound

Our robust bound consist of two parts:

$$M_\lambda \leq M^* = \min \left(\left\lfloor \frac{n_y - n_b + \mathbb{I}(y < b) - 1}{2} \right\rfloor, M_h \right),$$

Classification Bound: direct calculation;

$$\text{where } M_h = \max M, \text{ s.t. } n_l^\gamma - n_{h_M}^\gamma + \mathbb{I}(l < h_M) > 2M.$$

Explanation bound: needs to be examined through a numeration;

Proof: please check in paper and appendix.

This robust bound is deterministic against **arbitrary attack**, including random attack or **any adaptive attack**.

Details of the proofs and more ablation studies could be found in our full paper on the virtual page.