GRACED: A Plug-and-Play Solution for Certifiable Graph Classification

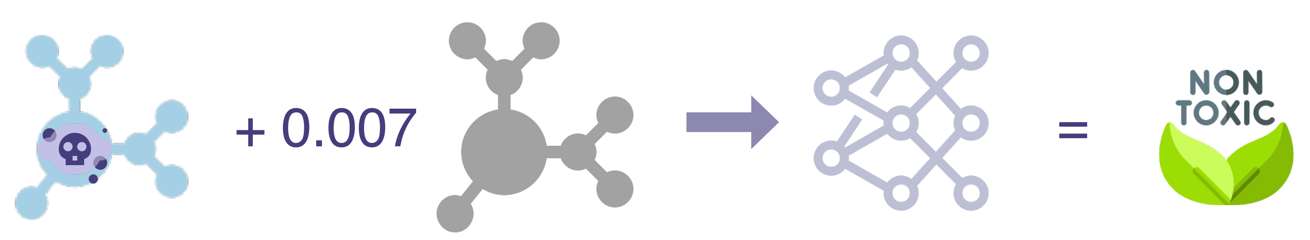
Xiaoyu Liang[1], Haohua Du[1], He Lu[1], and Fei Shang[2]

1. Beihang University, China, 2. University of Science and Technology of China, China

Adversarial Robustness for GNN

Adversarial Attack on Graph Classification

Graph Neural Network (GNN)-based Graph classifiers have been shown to be vulnerable to subtle modifications of the graph, which compromises its robustness when applied to tasks such as protein property analysis.



Certifiable Robustness with Randomized Smoothing

Certificate: Given the input graph $G = (X, A)$, base classifier f_θ and attack budget Δ, guarantee that for all $\delta \in \Delta$, $f_\theta(G + \delta) = f_\theta(G)$.

Randomized Smoothing (Cohen 2019, Bojchevski 2020) predicts the label with a smoothed base classifier $g_\theta(G)$:

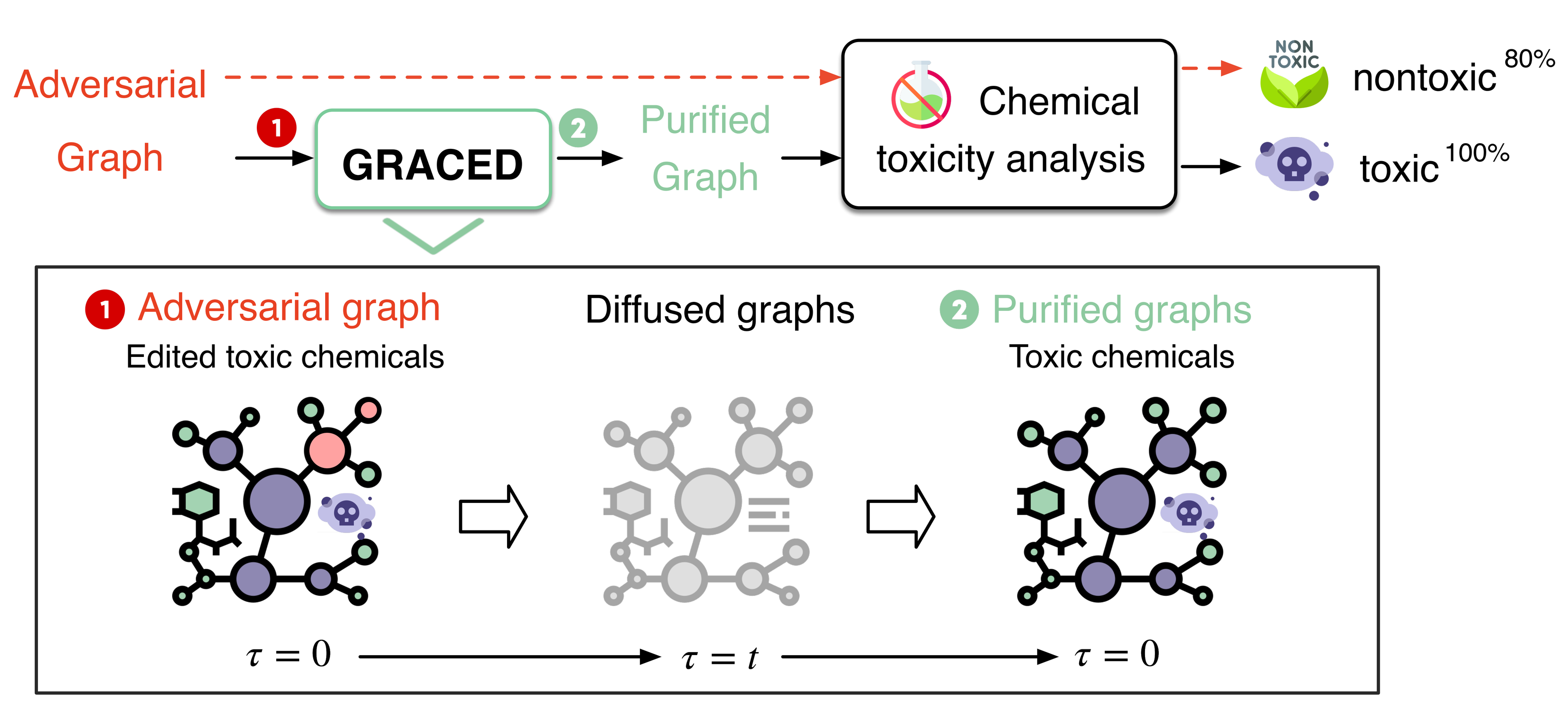
$$g_\theta(G) := \arg_y \max \Pr_{\tilde{G} \sim \phi(G)} [f_{\theta^*}(\tilde{G}) = y]$$

RS require retraining or fine-tuning on noisy samples
i) Retraining for diverse adversaries; ii) accuracy drop

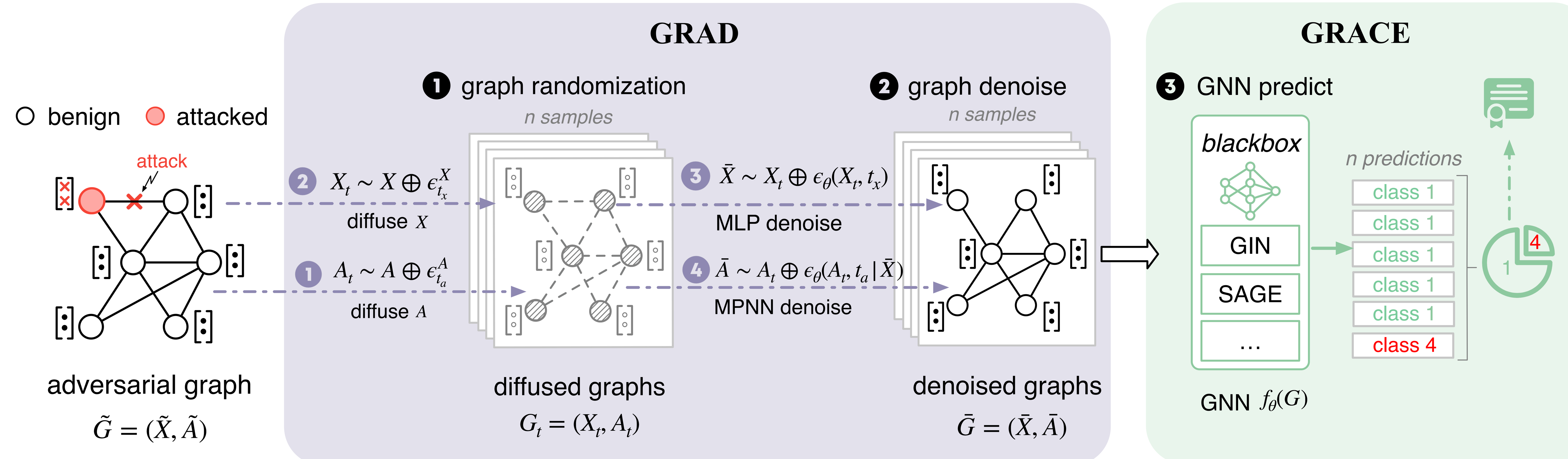
How to provide plug-and-play certified defense for GNN?

$$f_{\theta^*}(\tilde{G}) := f_\theta(\mathcal{D}(\tilde{G}))$$

Discrete Denoising Diffusion Probabilistic Models



GRACED: A Certified Graph Classification Solution



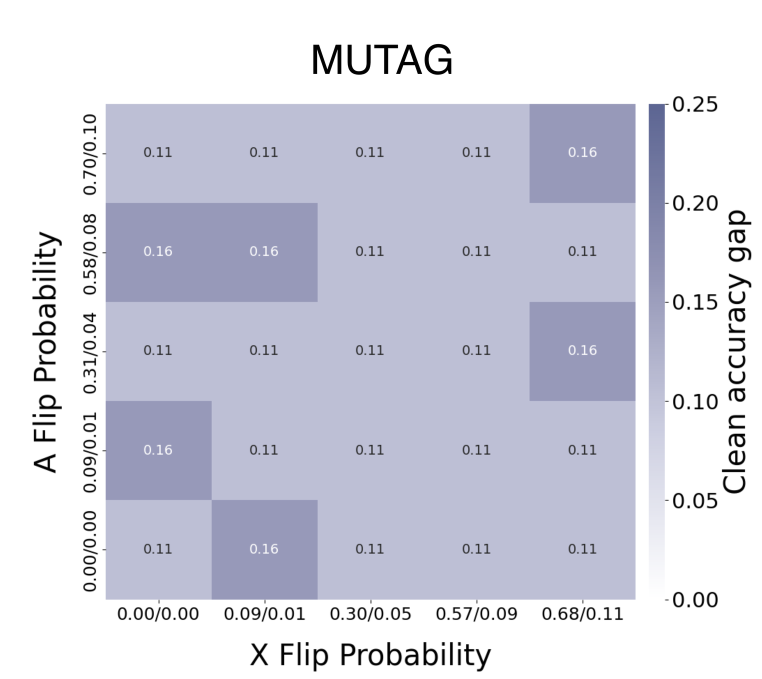
We propose a novel defense method—GRACED, which provides theoretical guarantees for the accuracy and robustness of graph classification without requiring knowledge of the attacker's capabilities or the classification model. The key idea behind our method is to leverage the denoising ability of feature diffusion models for adversarial data purification. We then demonstrate that this randomized purification approach can ensure certified robustness under specific attack budgets.

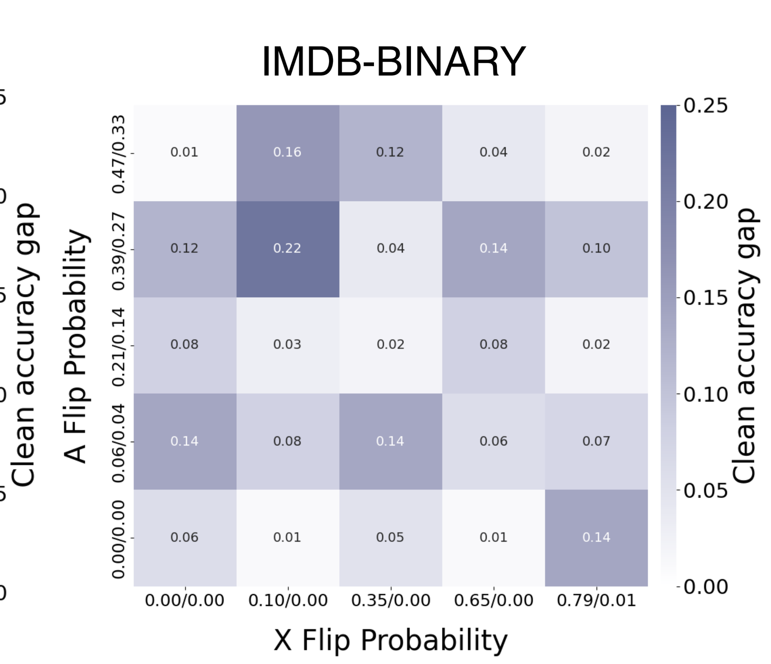
Evaluation Result

Extensive experiments show that graph classifiers using GRACED significantly outperform state-of-the-art classifiers. For instance, the accuracy on MUTAG improved by 11%, and the best results on IMDB showed a 14% increase.

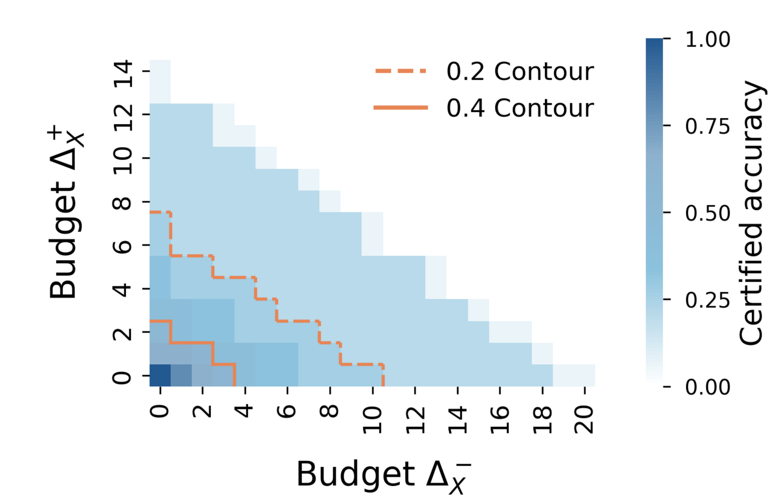
Clean Accuracy

Method	MUTAG	NCI1	PROTEINS	IMDB
Sparse	0.68	0.60	0.55	0.49
Hier.	0.52	**0.64**	0.63	0.48
Ber.	0.74	0.55	**0.67**	0.51
GRACED	**0.79**	**0.64**	**0.67**	**0.63**

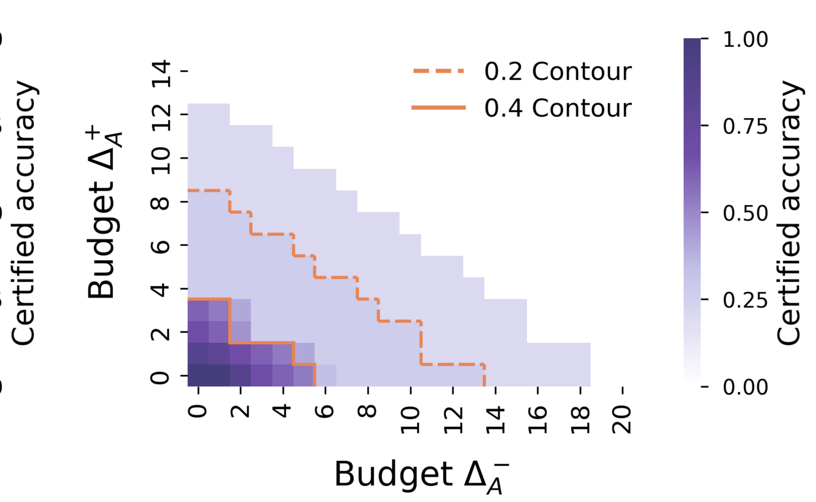




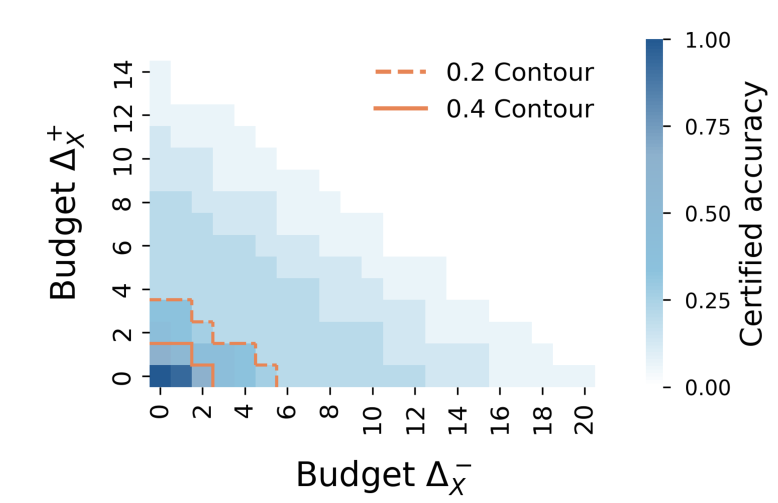
Certified Accuracy



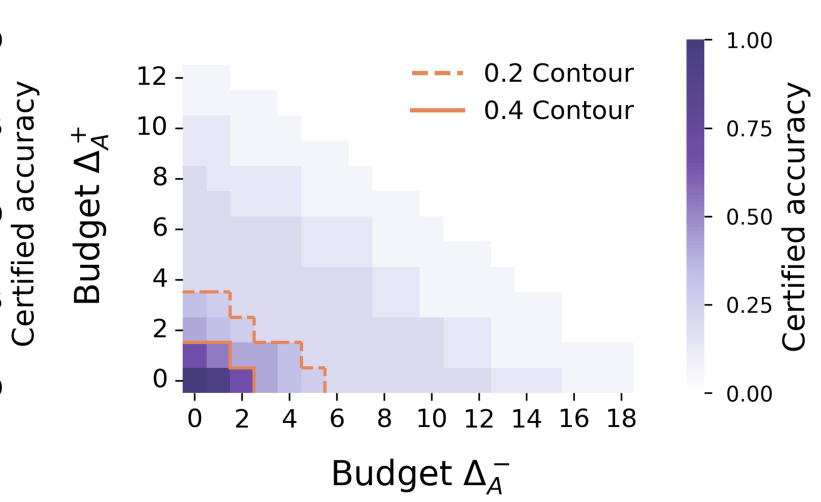
Singular certificate for attribute



Singular certificate for adjacent



Joint certificate with $\Delta_A = (0,0)$



Joint certificate with $\Delta_X = (0,0)$

[1] Bojchevski *et al.*, "Efficient Robustness Certificates for Discrete Data: Sparsity-Aware Randomized Smoothing for Graphs, Images and More," ICML 2020.
[2] Austin *et al.*, "Structured Denoising Diffusion Models in Discrete State-Spaces," NeurIPS 2021.
[3] Vignac *et al.*, "DiGress: Discrete Denoising diffusion for graph generation,", ICLR 2023.
[4] Scholten *et al.*, "Hierarchical Randomized Smoothing," NeurIPS 2023