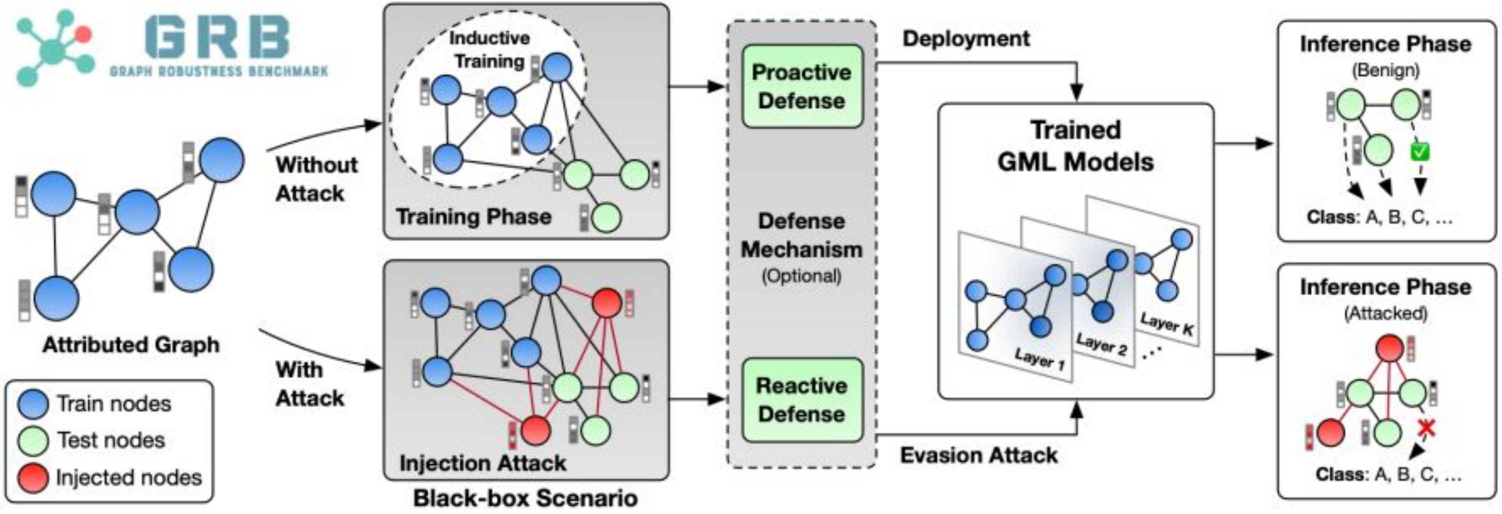
ICLR 24: Mitigating Emergent Robustness Degradation while Scaling Graph Learning

Xiangchi Yuan*, *Chunhui Zhang*, Yijun Tian, Yanfang Ye, Chuxu Zhang*

ICLR 24: Mitigating Emergent Robustness Degradation while Scaling Graph Learning

Xiangchi Yuan, *Chunhui Zhang, Yijun Tian, Yanfang Ye, Chuxu Zhang*

Problem: Defense/robust learning against graph adversarial attacks



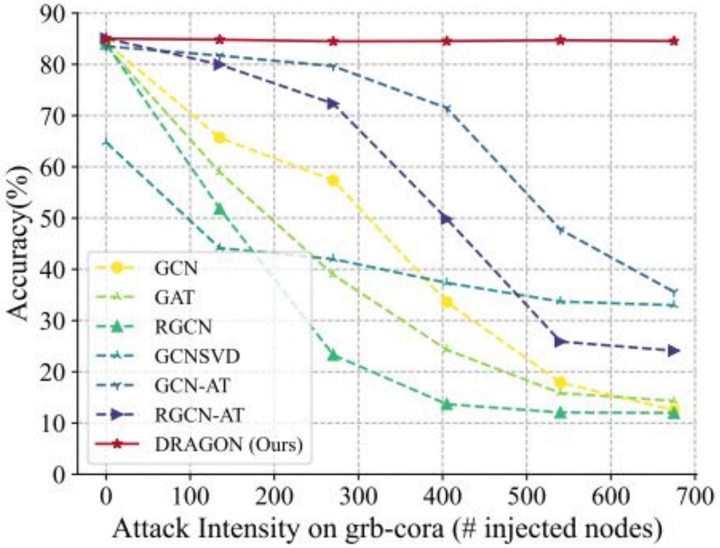
Challenges

Challenge 1. Severe robustness degradation:

when attack intensity surpasses a threshold of 300 injected nodes, error rates for many models surge by more than 50%.

Challenge 2. Scalability:

Many robust methods such as GNNGurad, SVD face scalability issue.



Overall Framework

Overall Framework:

Denoise → Robust classifier

Both two modules contributes to

Solving challenges.

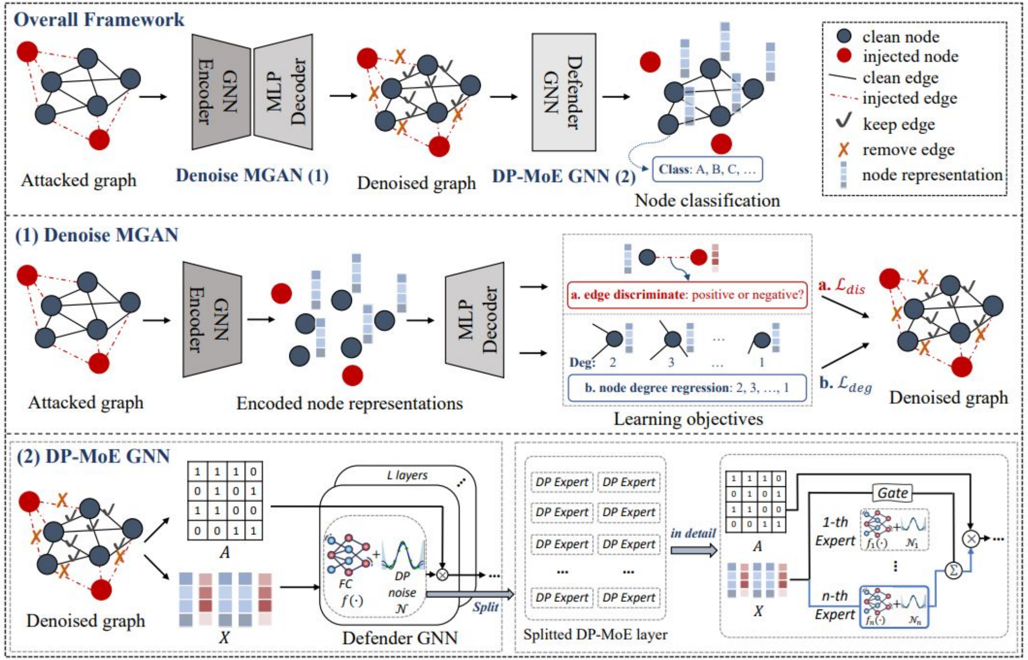
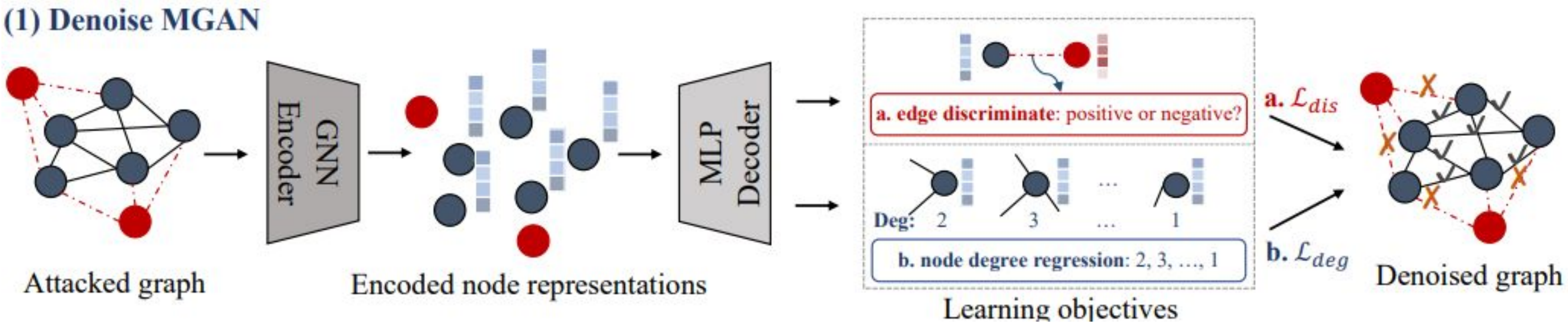


Figure 2: Our framework. First, (1) in Denoise MGAN, a cleaner graph is recovered by removing the edges connected to injected nodes, preventing their message-passing interactions with clean nodes. Second, the cleaner graph is classified using (2) in DPMoE GNN, which consists of a DP graph convolutional layer split into multiple DP expert networks with adjustable noise coefficients to handle attacks of different intensities.

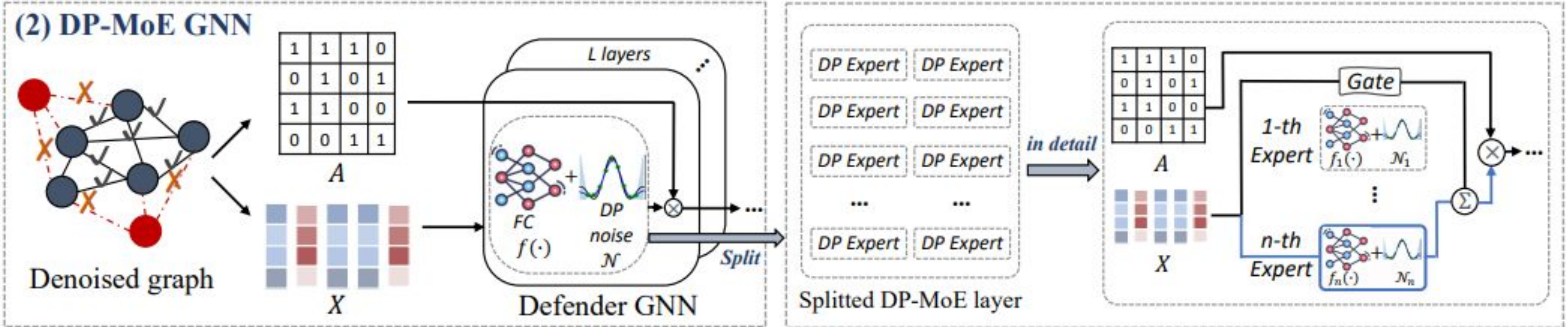
Contribution 1: Denoise module



Motivation: MAE (K. He, CVPR 2023) finds, reconstructed examples, although distinct from the ground truth, remained semantically plausible.

How: Mask paths – Self-supervise--Reconstruction

Contribution 2: Robust classifier



Differential Privacy and Robustness Connection: Perturbation of input will has bounded output.

Lemma 1. ***Robustness Guarantee for DPMoE.*** *For a GNN $f(\cdot)$ containing DPMoE which utilizes Gaussian DP, assume this mechanism lets the model output satisfy (σ, δ)-DP. If the expected value $\mathbb{E}$ of the model output satisfies the following property:*

$$\mathbb{E}(f_k(h_v^{(l)})) > e^{2\epsilon} \max_{i:i \neq k} \mathbb{E}(f_i(h_v^{(l)})) + (1 + e^{\epsilon})\delta, \tag{9}$$

then the label probability output vector $p(h_v^{(l)}) = (\mathbb{E}(f_1(h_v^{(l)})), \ldots, \mathbb{E}(f_K(h_v^{(l)})))$ of $f(\cdot)$ for node v satisfies the robustness: $\mathbb{E}(f_k(h_v^{(l)})) \geq \max_{i:i \neq k} \mathbb{E}(f_i(h_v^{(l)}))$.

Mitigating Severe Robustness Degradation on Graphs

Empirical finding: Matching DP noise magnitudes with different intensities can help model better defense attacks.

Mixture-of-Experts: MoE can select the most matching DP expert to handle the attack with specific intensity.

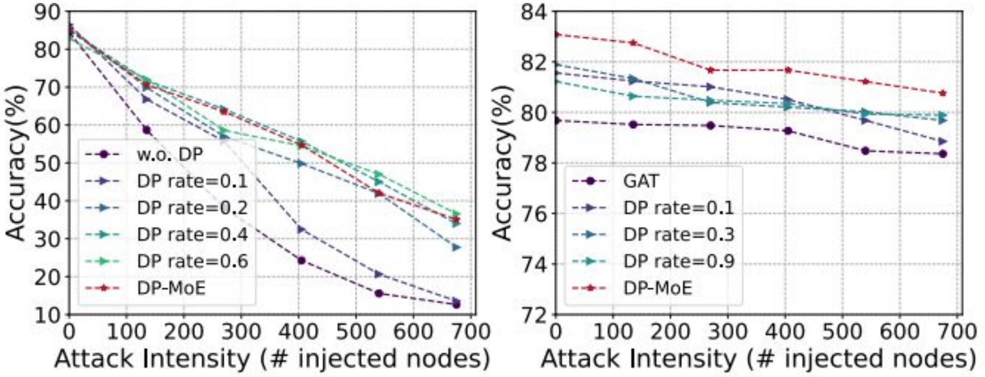


Figure 4: Different DP rates (scaling coefficient) on DRAGON w. single DP rate and w. multiple DP rates via DPMoE using standard training (left) and adversarial training (right) on *Cora* dataset.

Experiment results on datasets with different scales

Solve the challenges:
Anti-degraded robustness and scalability

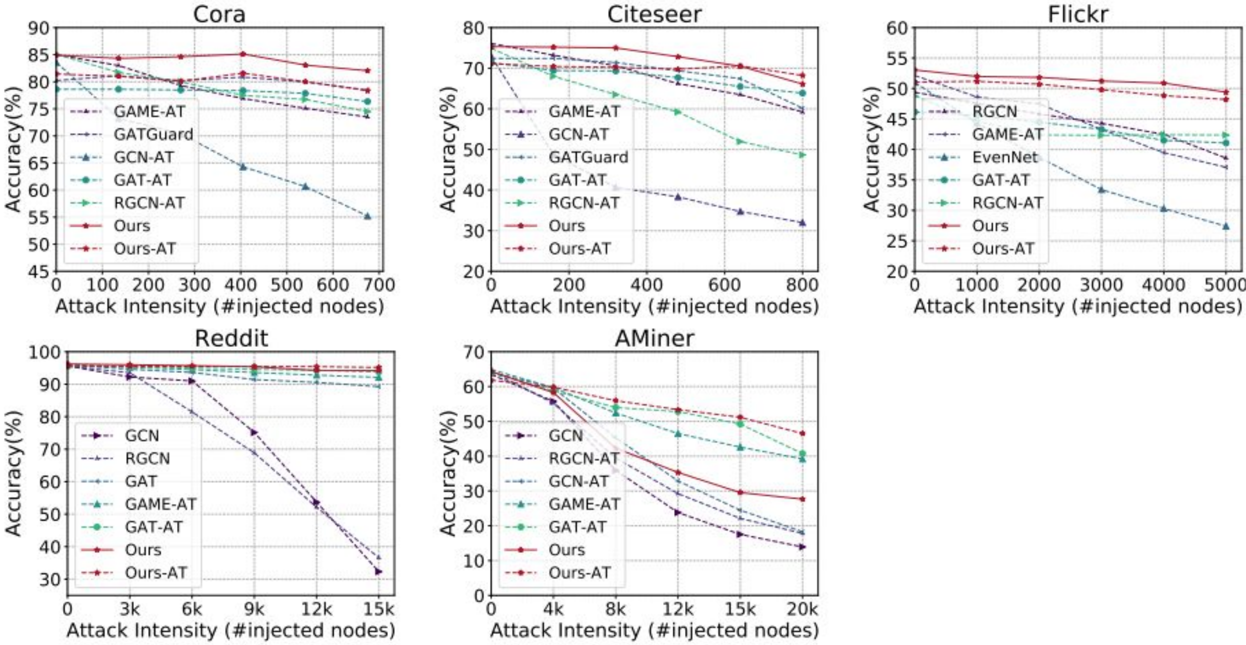


Figure 8: The performance of top-5 baselines and our method under the HAO Attack.