GNNGuard: Defending Graph Neural Networks against Adversarial Attacks

Xiang Zhang, xiang_zhang@hms.harvard.edu
Marinka Zitnik, marinka@hms.harvard.edu

HARVARD UNIVERSITY
NEURAL INFORMATION PROCESSING SYSTEMS

1. Take-Home Message

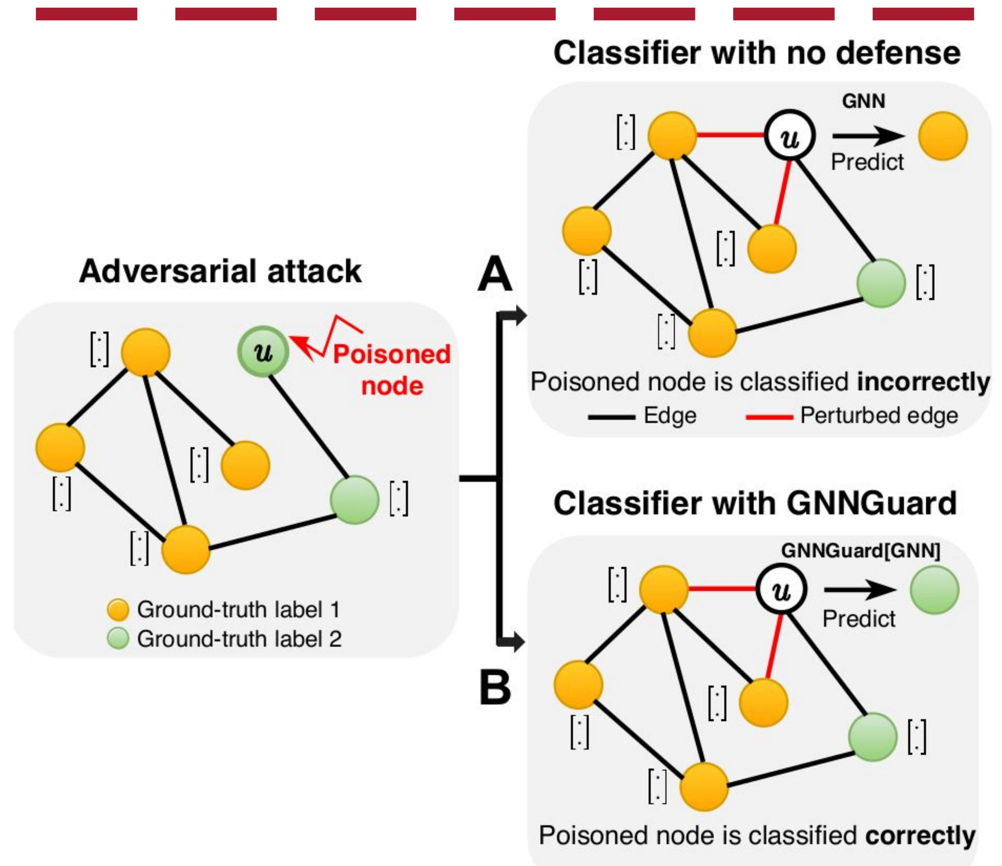
GNNGuard is a model-agnostic approach that can defend any Graph Neural Network against a variety of poisoning adversarial attacks.

2. Featured Properties

- **Defense against a variety of attacks:** e.g., directly targeted, influence targeted, and non-targeted attacks
- **Integrates with any GNNs**
- **State-of-the-art performance on clean graphs**
- **Homophily and heterophily graphs:** the first technique defending GNNs against attacks on both homophily and heterophily graphs

3. Motivation

- GNNs are highly vulnerable to adversarial attacks
  - Adversarial attacks: inject carefully-designed perturbations (e.g., fake edges) to graph to degrade GNN classifier
- The vulnerability significantly prevent GNNs from real-world applications

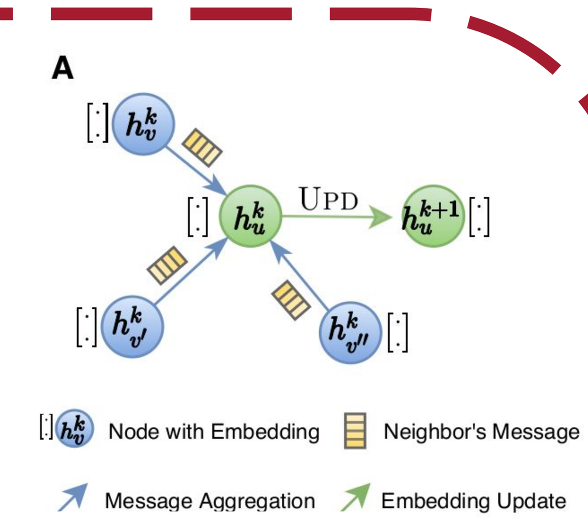


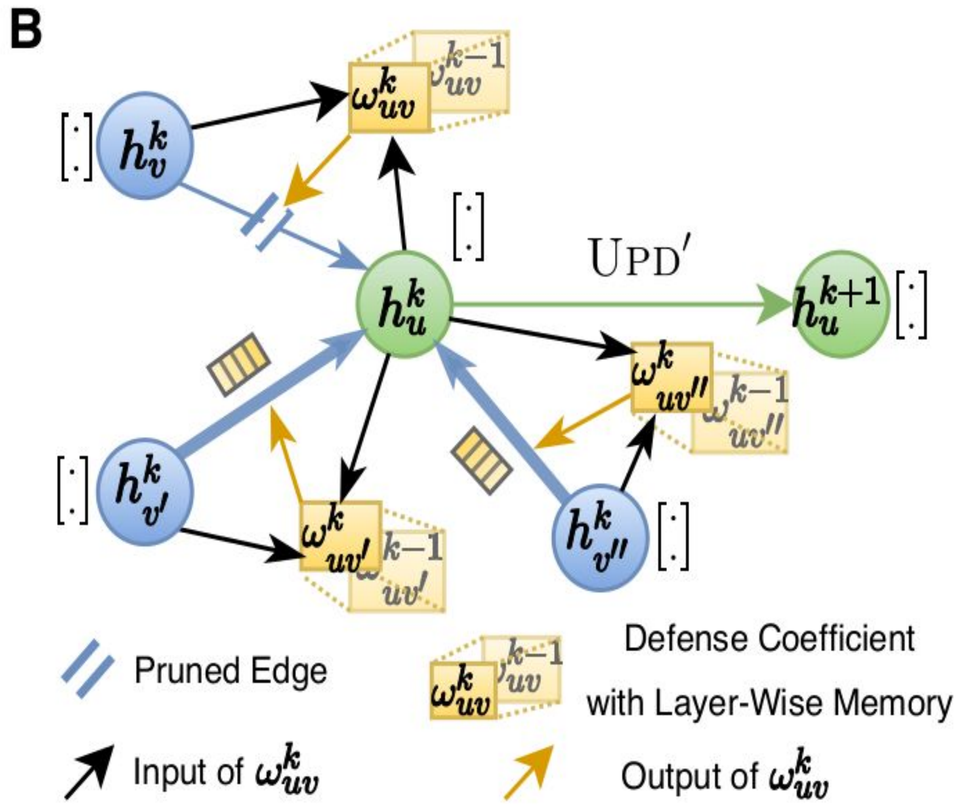
- Panel A (without GNNGuard): Missclassification
- Panel B (**with GNNGuard**): **correct classification**

4. Method

GNNGuard detects fake edges and alleviate the negative impact on prediction by removing them or assigning them lower weights in neural message passing.

Compared to a typical GNN (panel A), GNNGuard (panel B) controls the message stream, such as blocking the message from irrelevant neighbors while strengthening messages from highly-related ones.

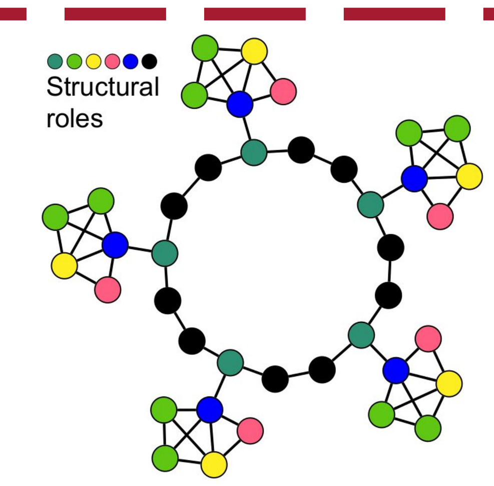




GNNGuard contains two key components:

- **Neighbor Importance Estimation:** 1) estimate the importance of each edge in neighborhood; 2) prune fake edges and assign lower weights to likely-fake edges
- **Layer-Wise Graph Memory:** 1) keeps partial memory of the pruned graph structure from the previous layer; 2) smooth the evolution of edge pruning

GNNGuard can defend heterophily graph against adversarial attack by estimating neighbor importance through graphlet signature.



5. Experiments

GNNGuard outperforms existing defense approaches by **15.3%** on average across five GNNs, three cutting-edge defense baselines, and three adversarial attackers.

Dataset Description

Dataset	N	E	M	C	Node features
Cora	2,485	5,069	1,433	7	Binary
Citeseer	2,110	3,668	3,703	6	Binary
ogbn-arxiv	31,971	71,669	128	40	Continuous
DP	22,552	342,353	73	519	Continuous
Synthesized	1,000	3,200	-	6	-

Results in Graphs with Homophily

Model	Dataset	No Attack	Attack	GNN-Jaccard	RobustGCN	GNN-SVD	GNNGuard
GCN	Cora	0.826	0.250	0.525	0.215	0.475	**0.705**
	Citeseer	0.721	0.175	0.435	0.230	0.615	**0.720**
	ogbn-arxiv	0.667	0.235	0.305	0.245	0.370	**0.425**
	DP	0.682	0.215	0.340	0.315	0.395	**0.430**
GAT	Cora	0.827	0.245	0.295	0.215	0.365	**0.625**
	Citeseer	0.718	0.265	0.575	0.230	0.575	**0.765**
	ogbn-arxiv	0.669	0.210	0.355	0.245	0.445	**0.520**
	DP	0.714	0.205	0.320	0.315	0.335	**0.445**
GIN	Cora	0.831	0.270	0.375	0.215	0.375	**0.645**
	Citeseer	0.725	0.285	0.570	0.230	0.570	**0.755**
	ogbn-arxiv	0.661	0.315	0.425	0.245	0.475	**0.640**
	DP	0.719	0.245	0.410	0.315	0.405	**0.460**
JK-Net	Cora	0.834	0.305	0.445	0.215	0.425	**0.690**
	Citeseer	0.724	0.275	0.615	0.230	0.610	**0.775**
	ogbn-arxiv	0.678	0.335	0.375	0.245	0.325	**0.635**
	DP	0.726	0.220	0.335	0.315	0.360	**0.450**
Graph SAINT	Cora	0.821	0.225	0.535	0.235	0.460	**0.695**
	Citeseer	0.716	0.195	0.470	0.350	0.395	**0.770**
	ogbn-arxiv	0.683	0.245	0.365	0.245	0.315	**0.375**
	DP	0.739	0.205	0.315	0.295	0.330	**0.485**

Results in Graphs with Heterophily

Model	No Attack	Attack	GNN-Jaccard	RobustGCN	GNN-SVD	GNNGuard
GCN	0.834	0.385	N/A	0.525	0.595	**0.715**
GAT	0.851	0.325	N/A	0.575	0.635	**0.770**
GIN	0.891	0.450	N/A	0.575	0.650	**0.775**
JK-Net	0.889	0.425	N/A	0.575	0.640	**0.735**
GraphSAINT	0.876	0.415	N/A	0.575	0.625	**0.755**

HARVARD MEDICAL SCHOOL
HDSI Harvard Data Science Initiative
BROAD INSTITUTE

Project website