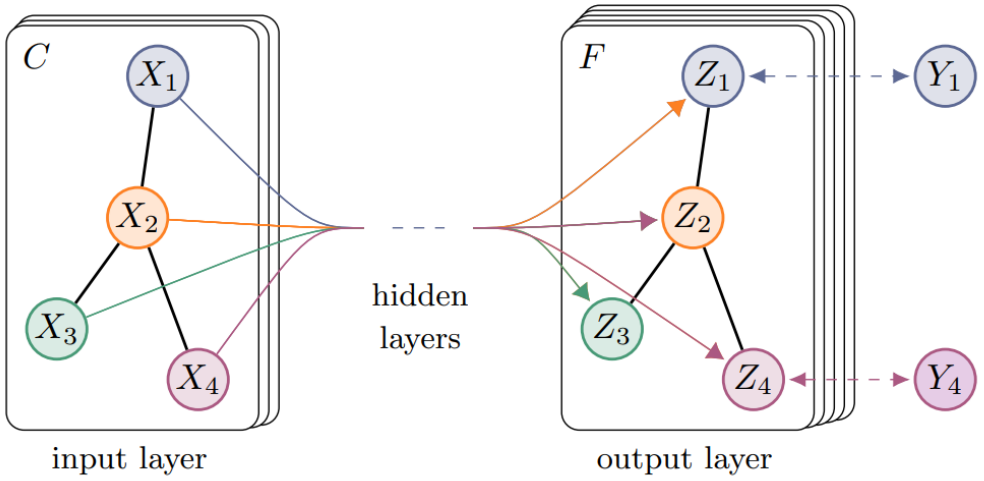
Provably Robust Node Classification via Low-Pass Message Passing

Yiwei Wang, Shenghua Liu, Minji Yoon, Hemank Lamba, Wei Wang, Christos Faloutsos, Bryan Hooi

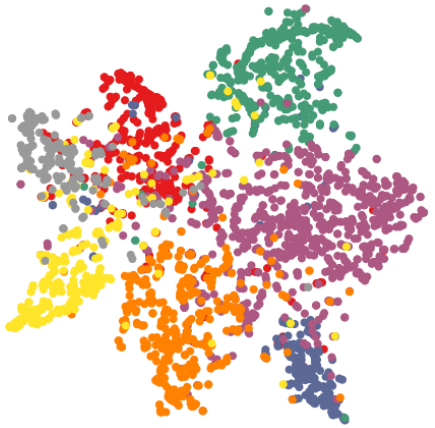
21th Oct 2020

Graph Convolutional Networks

- GCNs have achieved state-of-the-art performance on node classification
- GCNs follow Message Passing mechanism to make prediction
  - Aggregate semantic representations of each node and its neighbors at each layer
  - Give similar predictions to the connected nodes
- However, recent works have shown that GCNs are vulnerable to adversarial attacks, such as additions or deletions of adversarially-chosen edges in the graph

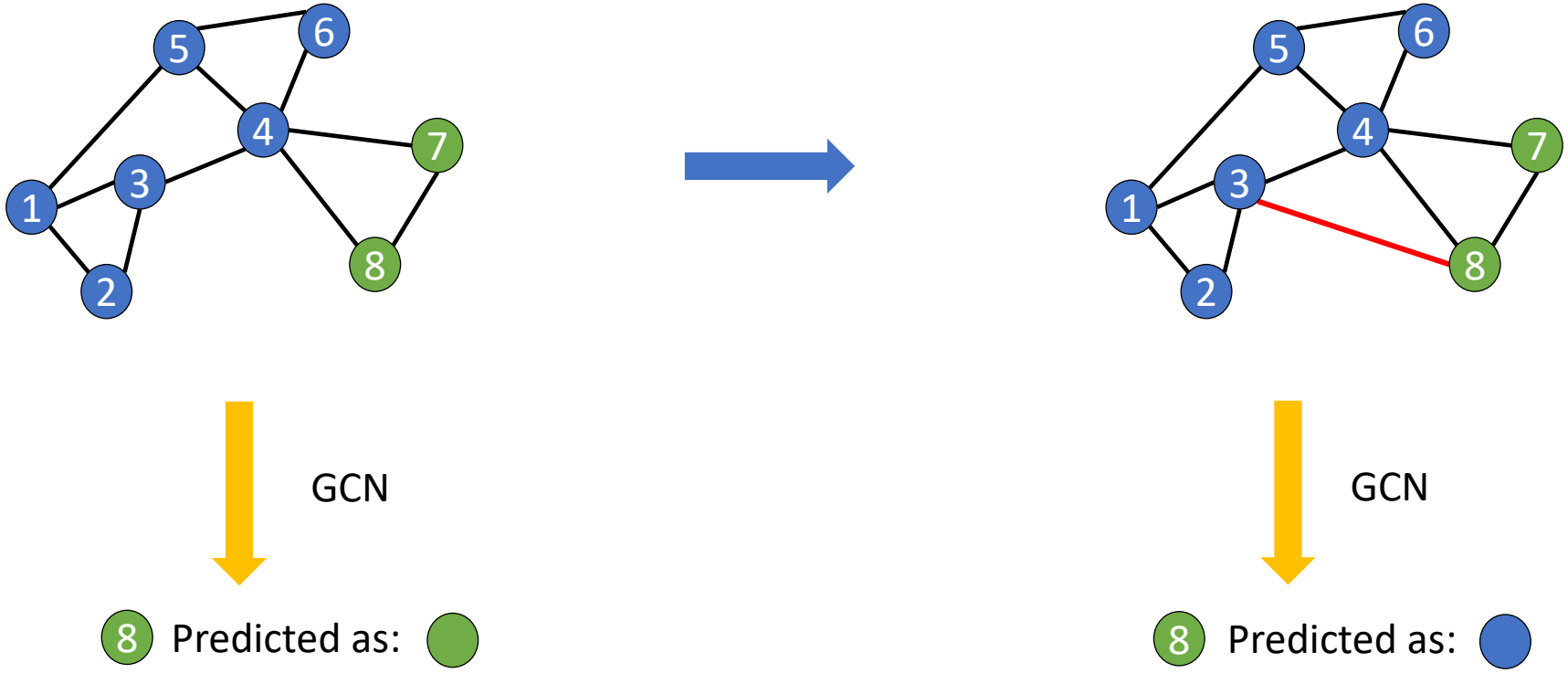


(a) Graph Convolutional Network

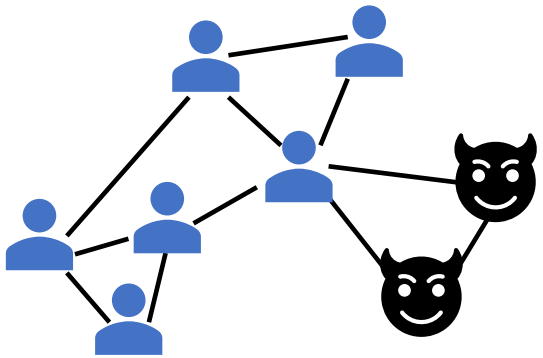


(b) Hidden layer activations

An Example of Adversarial Attacks on GCNs

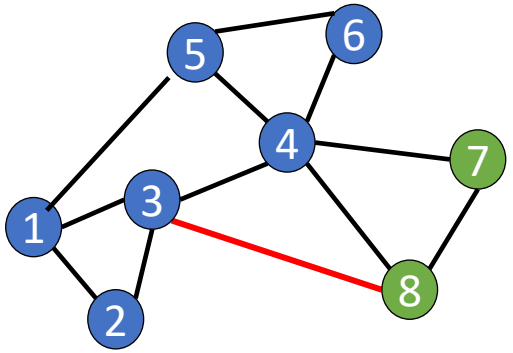


Consequences



- Financial Systems
  - Credit Card Fraud Detection
- Recommender Systems
  - Social Recommendation
  - Product Recommendation
- ….

Problem Statement & Our Target



- The adversary adds or removes any edges for the targeted graph nodes in order to affect their classification output as much as possible, subject to a set of budget constraints
  - The budget of the adversary is measured in terms of the fraction of each node's edges that the adversary can modify.
- Our targets
  - We propose a robust node classification method that effectively defends against graph structural attacks.
  - Additionally, we focus on obtaining an upper bound for the deviations of the log probability of any node and any class under perturbations.

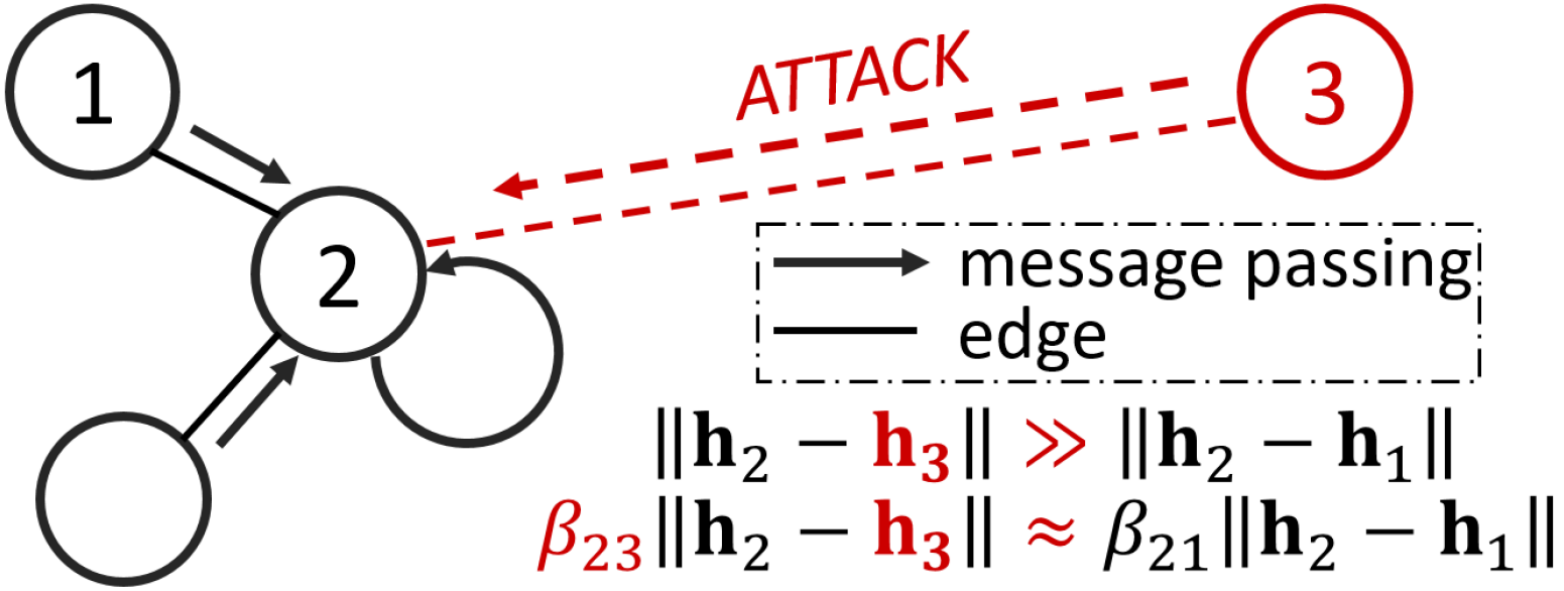
Original Message Passing

- The message passing strength from different neighbors (edge weight) are equal
- If the perturbation $\mathbf{h}_j - \mathbf{h}_i$ is very large, it will be propagated to the aggregated vector $\mathbf{h}_i^{aggre}$

$$\mathbf{h}_i^{aggre} = \sum_{j \in \mathcal{N}_i} \frac{1}{d_i} \mathbf{h}_j = \mathbf{h}_i + \frac{1}{d_i} \sum_{j \in \mathcal{N}_i \setminus \{i\}} (\mathbf{h}_j - \mathbf{h}_i)$$

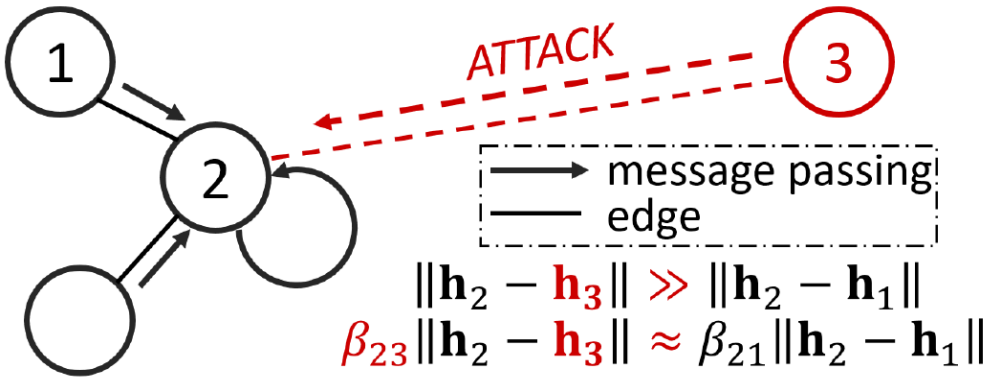
Our Low Pass Message Passing

- Our low-pass edge weights inhibit the effects from adversarial edges
  - The edge between nodes 2 and 3 is injected for attacking node 2
  - With the original message passing, the feature of node 3 affects 2 heavily. But our low-pass edge weight β_{23} inhibits this effect



Our Low-Pass Message Passing

- We aim to limit the influence that a node can have on another
- Denote $R > 0$ as the threshold for controlling our low-pass message passing
- Gradually reducing the weight as the distance between them exceeds R

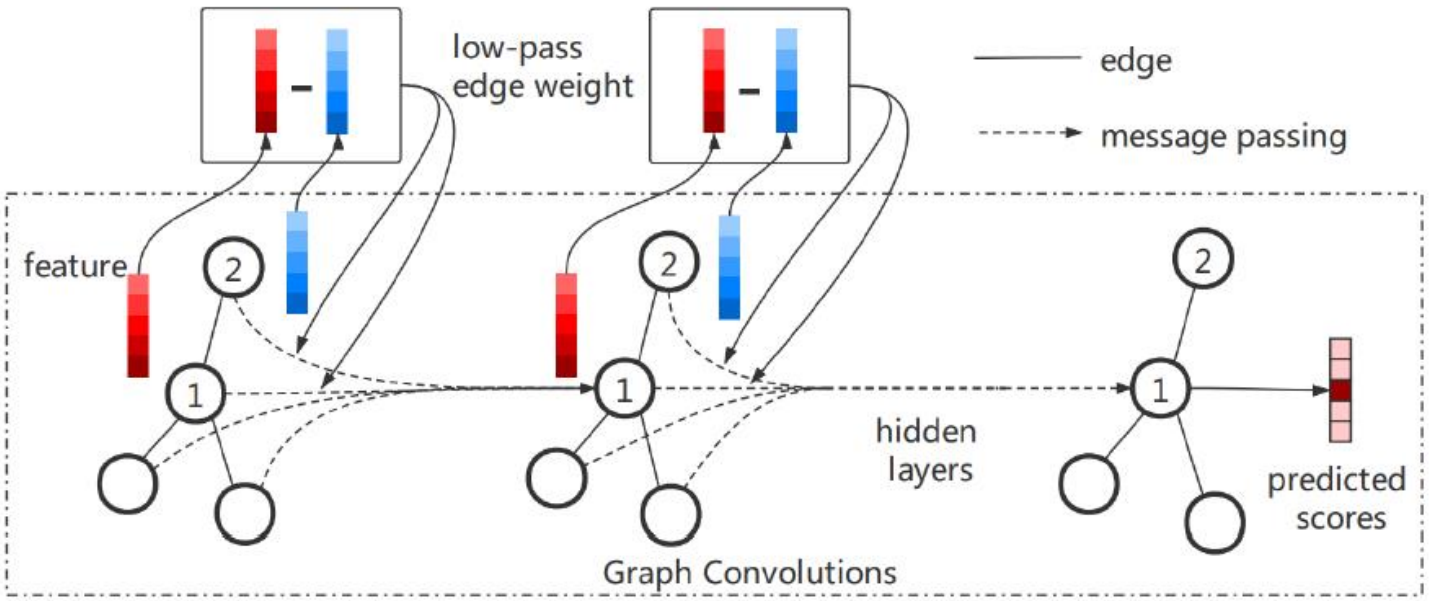


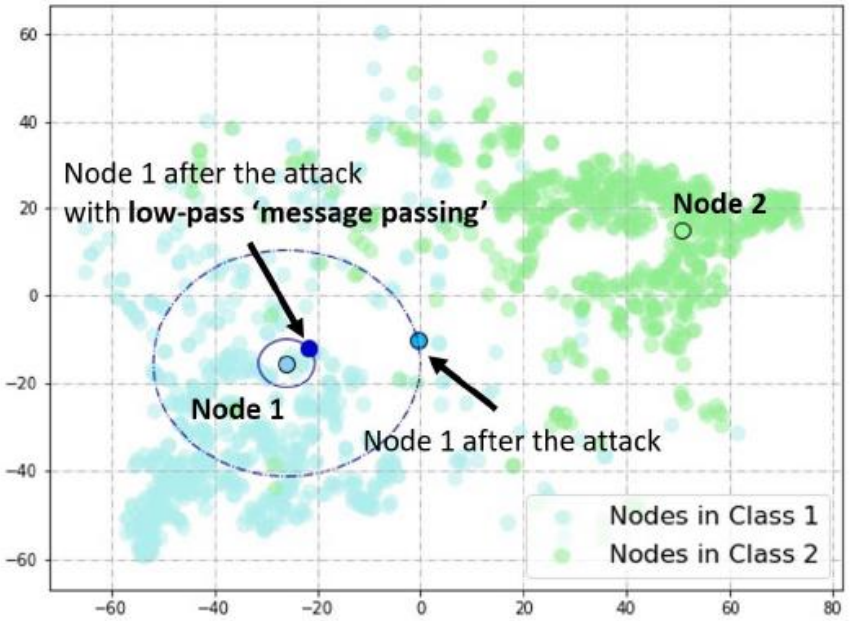
$$\beta_{ij}^{(l)} = \frac{R}{\max(R, \|\mathbf{h}_i^{(l)} - \mathbf{h}_j^{(l)}\|)}$$

$$B_{ij}^{(l)} = \begin{cases} \beta_{ij}^{(l)}/d_i & \text{if } j \in \mathcal{N}_i \setminus \{i\} \\ 1 - \sum_{j \in \mathcal{N}_i \setminus \{i\}} \beta_{ij}^{(l)}/d_i & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

Our Low-Pass Message Passing

- Framework
- Case study
  - The final-layer hidden representations of the nodes retrieved by GCN belonging to two classes in Citeseer (visualized via t-SNE).
  - After the adversarial injection of edge between nodes 1 and 2, severe deviations happen on nodes 1. But with our low-pass ‘message passing’, its deviation is inhibited.





Provable Robustness is Essential

Provable Robustness

Lemma 1. *Given any matrices* $\mathbf{X}, \mathbf{Y}$, *we have*

$$\|\mathbf{X} - \mathbf{Y}\| \geq \|\sigma(\mathbf{X}) - \sigma(\mathbf{Y})\|$$

Provable Robustness

Lemma 2. *Given the budget of the graph structural attacks as* $\mathcal{M}_i/d_i \leq \Delta < 1, \forall i$, *the deviation of* $\mathbf{H}^{(l+1)}$ *is bounded by:*

$$\|\mathbf{H}^{(l+1)} - \check{\mathbf{H}}^{(l+1)}\| \leq \|\mathbf{W}^{(l)}\|\left(\frac{2R\Delta}{1+\Delta} + 4\|\mathbf{H}^{(l)} - \check{\mathbf{H}}^{(l)}\|\right)$$

Provable Robustness

Theorem 1. *Given the attack budget:* $\mathcal{M}_i/d_i \leq \Delta < 1$, *we have:*

$$\|\mathbf{H}^{(L)} - \check{\mathbf{H}}^{(L)}\| \leq 2R\frac{\Delta}{1+\Delta}\left(\sum_{s=0}^{L-1}\|\mathbf{W}^{(s)}\|\prod_{l=s+1}^{L-1}4\|\mathbf{W}^{(l)}\|\right).$$

Provable Robustness

Theorem 2. *The adversarial perturbation in the log-probability for node i belonging to any class c is bounded by:*

$$|\log \check{P}_{ic} - \log P_{ic}| \leq \|\mathbf{H}^{(L)} - \check{\mathbf{H}}^{(L)}\|$$
$$\leq 2R\frac{\Delta}{1+\Delta}\left(\sum_{s=0}^{L-1}\|\mathbf{W}^{(s)}\|\prod_{l=s+1}^{L-1}4\|\mathbf{W}^{(l)}\|\right) \quad (8)$$

Experiments

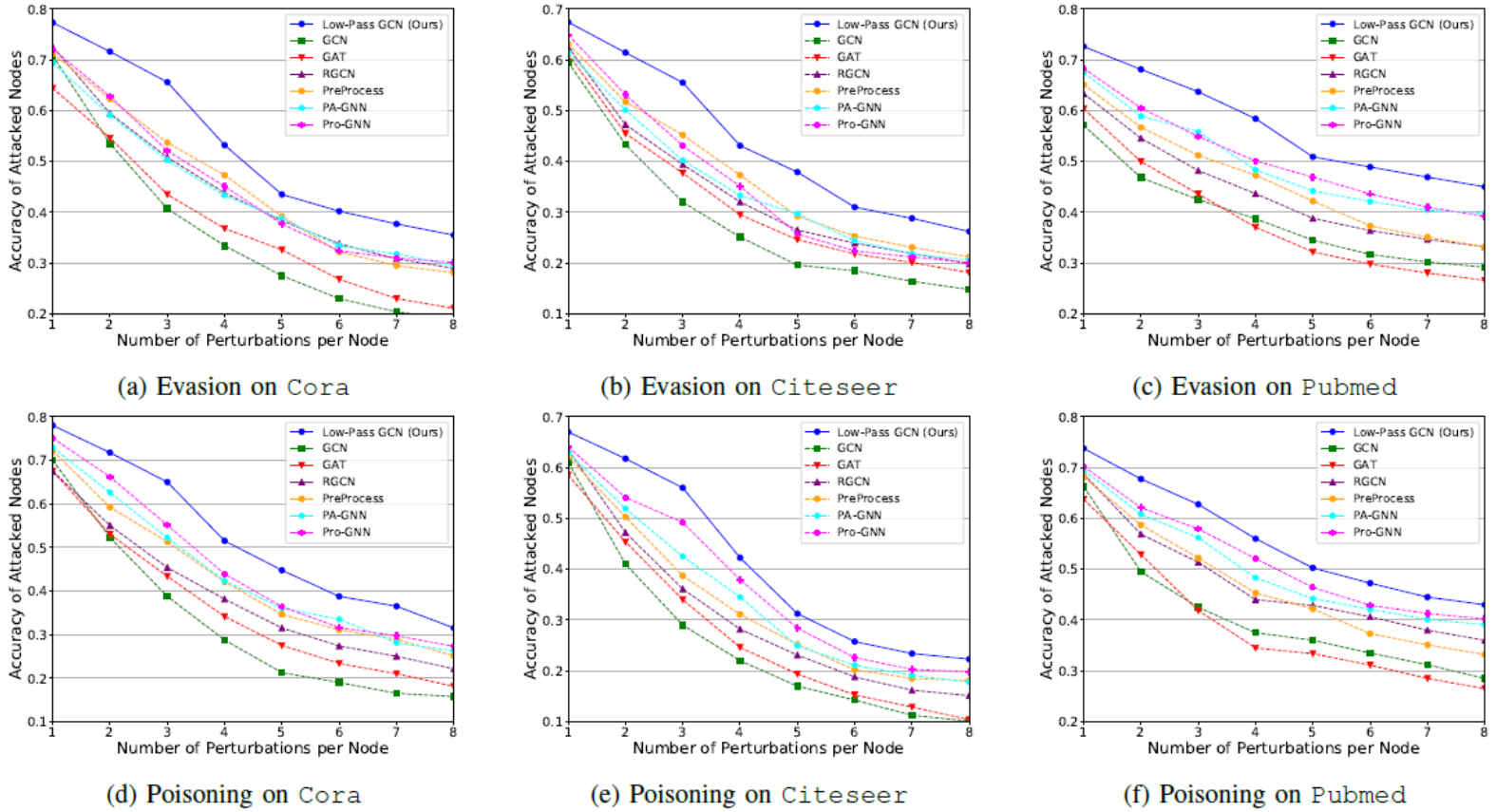
Dataset	# Nodes	# Edges	# Classes	# Features
Cora	2,708	5,429	7	1,433
Citeseer	3,327	4,732	6	3,703
Pubmed	19,717	44,338	3	500
Co-CS	18,333	81,894	67	6,805
Co-Phy	34,493	247,962	5	8,415

Experiments

Table II: Results of node classification with the random splits of 20 labeled examples per class, in terms of test accuracy (%). We report mean and standard derivations of 2000 trials on 20 splits with randomly selected targeted nodes.

Attack Method	Defense Method	Cora	Citeseer	Pubmed	Co-CS	Co-Phy
RL-S2V	GCN [3]	51.1±1.2	42.1±1.4	59.8±1.3	62.1±0.8	63.2±0.7
	JKNet [12]	52.7±1.4	44.1±1.1	61.1±0.7	61.7±1.2	62.8±1.5
	IncepGCN [13]	52.9±1.2	44.3±1.0	60.5±0.9	61.9±1.1	63.1±0.9
	GAT [29]	51.8±1.5	43.2±0.6	60.3±1.1	61.4±1.4	62.5±1.3
	RGCN [21]	58.6±1.3	52.8±1.4	66.5±1.6	68.3±1.9	69.5±2.1
	PreProcess [25]	61.3±2.3	57.2±1.2	69.8±1.4	72.9±1.5	73.1±0.9
	PA-GNN [30]	57.5±0.7	51.4±0.8	66.7±1.3	65.8±1.3	64.3±1.1
	Pro-GNN [31]	59.1±0.7	52.8±0.8	69.3±1.3	69.0±1.3	68.9±1.1
	Low-Pass GCN (Ours)	66.2±0.8	62.3±0.7	75.1±1.2	**77.1±0.7**	**79.2±0.9**
	Low-Pass JKNet (Ours)	66.7±0.7	62.5±0.9	**75.9±1.2**	76.9±0.6	79.0±1.2
	Low-Pass IncepGCN (Ours)	**66.9±0.7**	**62.8±0.9**	75.0±0.5	76.7±0.8	79.1±1.1
NETTACK	GCN [3]	49.2±1.4	44.8±1.2	60.7±1.1	61.1±1.1	61.4±0.7
	JKNet [12]	49.9±1.1	44.3±0.7	61.2±1.2	60.9±1.3	61.5±1.0
	IncepGCN [13]	49.8±1.1	45.2±1.3	60.9±0.7	61.0±0.8	61.1±1.1
	GAT [29]	49.3±1.3	44.2±1.1	60.2±0.5	60.3±1.7	61.2±1.1
	RGCN [21]	52.7±1.3	46.3±1.4	63.2±0.9	63.1±0.7	62.2±1.0
	PreProcess [25]	53.2±0.9	48.6±1.1	63.9±1.1	64.2±0.6	65.4±0.8
	PA-GNN [30]	57.9±1.1	52.1±1.3	68.4±1.4	68.4±1.1	67.6±1.5
	Pro-GNN [31]	61.2±1.4	56.2±0.8	71.9±1.2	73.7±0.9	72.1±1.1
	Low-Pass GCN (Ours)	64.3±0.8	60.2±1.1	74.3±0.9	**75.2±0.7**	**76.1±1.0**
	Low-Pass JKNet (Ours)	**64.8±0.6**	59.9±1.0	**74.5±0.8**	75.0±0.5	75.9±1.4
	Low-Pass IncepGCN (Ours)	64.5±0.5	**60.7±0.9**	74.2±1.3	75.0±0.8	75.4±1.3

Experiments



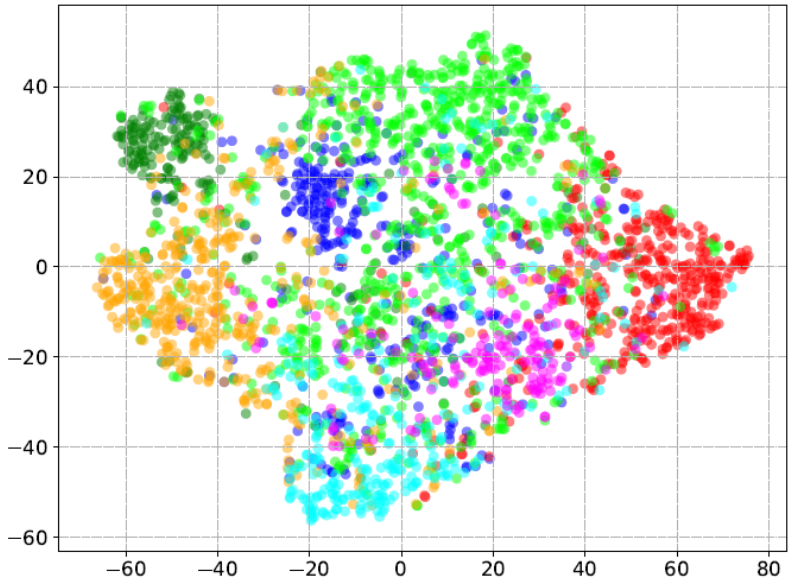
(a) Evasion on Cora (b) Evasion on Citeseer (c) Evasion on Pubmed

(d) Poisoning on Cora (e) Poisoning on Citeseer (f) Poisoning on Pubmed

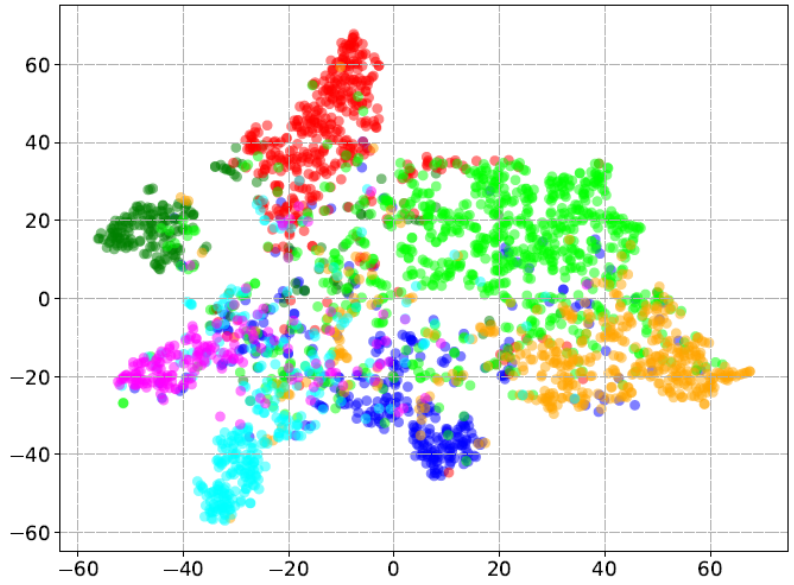
Figure 3: Classification accuracy vs. number of perturbations per node. Results of node classification with the random splits of 20 labeled examples per class are presented. We report the mean values of 2000 trials on 20 splits with randomly selected targeted nodes. We apply the Greedy method to attack the graph structures.

Experiments

- Low-Pass ‘message passing’ helps GCN to learn more robust representations
- We visualize the final-layer hidden representations of the all the nodes in the Cora dataset given by GCN and the GCN with low-pass ‘message passing’ by t-SNE.



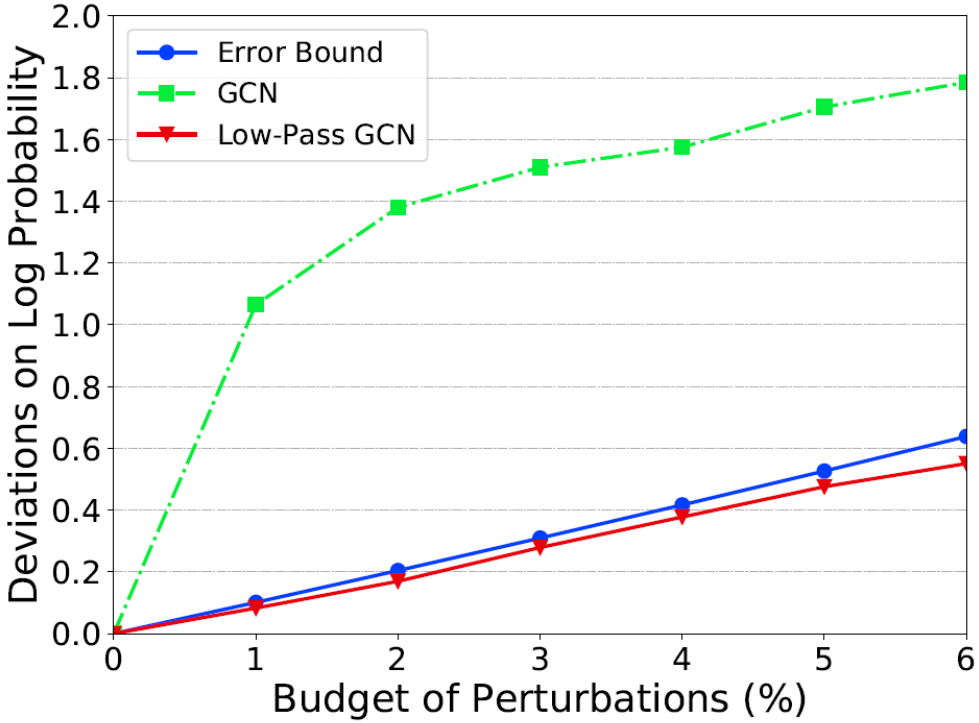
(a) GCN



(b) Low-Pass GCN

Experiments

- We provide an effective theoretical bound with our low-pass ‘message passing’
  - We plot largest element-wise deviations vs. adversarial attack budget
  - The blue line is the theoretical upper bound on the deviation of log probabilities derived for our proposed method.



Experiments

Table III: Node classification accuracy (%) of our low-pass 'message passing' with different R values. We report mean values of 2000 trials on 20 splits attacked by RL-S2V. The best value at each row is **boldface**.

Dataset	0.1	0.2	0.5	1	2	5	10
Cora	58.1	63.8	66.1	**66.2**	65.1	53.2	51.4
Citeseer	52.1	56.3	62.1	**62.3**	58.3	42.9	42.6
Pubmed	71.2	72.1	**75.3**	75.1	72.1	62.2	60.6
Co-CS	71.6	72.3	76.8	**77.1**	75.1	63.2	62.9
Co-Phy	72.1	73.5	79.1	**79.2**	77.4	63.5	63.2

Thank You!