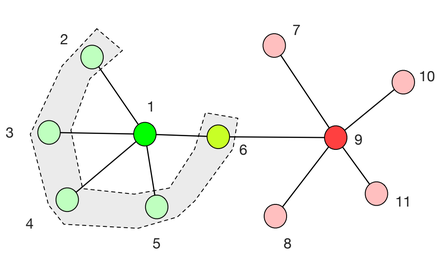
Detection and Defense of Topological Adversarial Attacks on Graphs

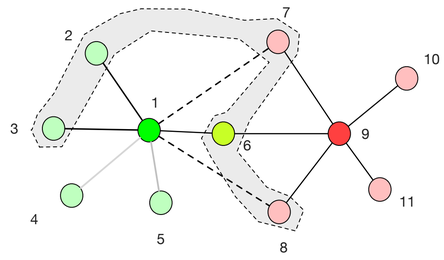
Yingxue Zhang, Florence Regol, Soumya Pal, Sakif Khan, Liheng Ma, Mark Coates
Huawei Noah's Ark Lab, McGill University

Introduction

- Robustness analysis is important for graph neural networks
- Some classical graph neural networks have been shown to be vulnerable to adversarial attacks [4], [5]
- Adversarial perturbations to graph structure can effectively induce classification errors



Clean local topology for node 1



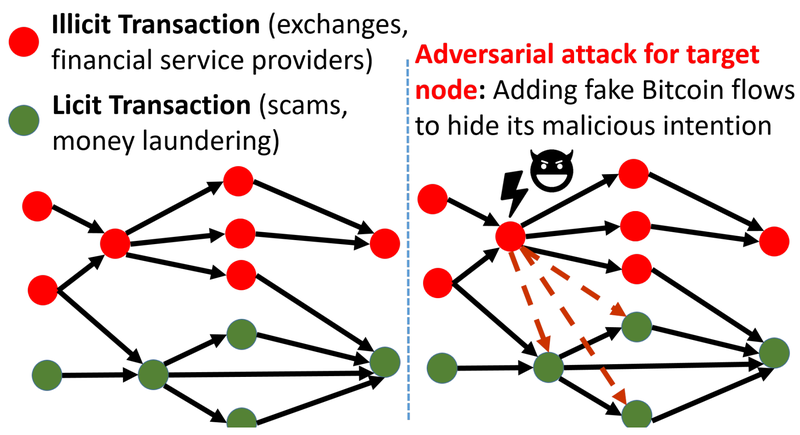
Perturbed local topology for node 1

- Recently proposed graph neural networks with structure exploration component demonstrate robustness to adversarial perturbations of the topology [2], [3]

Adversarial attack on graphs

- Random perturbation of targeted node: delete a portion of existing edges and add the same amount of neighbors with different labels
- Focused perturbation of targeted node (Nettack [4]): make "unnoticeable" perturbations which degrade classification margin as much as possible
- Global attack (Meta-Learning attack [5]): degrade overall classification performance while maintaining "unnoticeability" of local perturbations

A practical demo for adversarial attacks

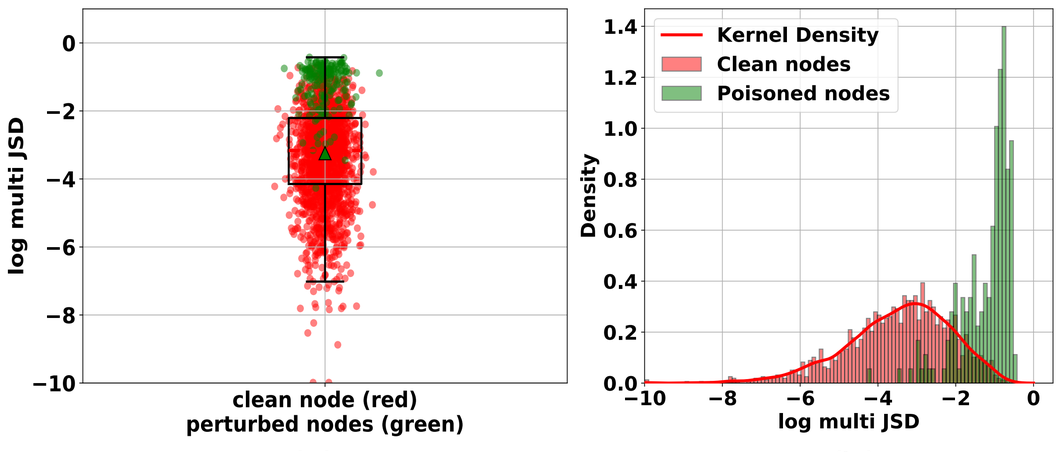


Adversarial attack detection on graphs

- Adversarial perturbations of structure create discrepancy between center node information of and those of its neighbours.
- We measure this discrepancy by multi-distribution Jensen-Shannon Divergence between a set of softmax probabilities.

$$\mathrm{JSD}\left(\mathcal{P}_i\right) = H\left(\frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} p_j\right) - \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} H\left(p_j\right) . \quad (1)$$

- Visualization



(a) (b)

Box plots of Multi-JSD statistics after log transform for unperturbed nodes (red) and perturbed nodes (green). (b) using kernel density function to fit the log-transformed statistics from the unperturbed nodes (Citeseer dataset, under Nettack). (use Citeseer dataset as a example)

Experimental Results

- **Detection comparison**
- Goal: Examine the effectiveness of the proposed detector.

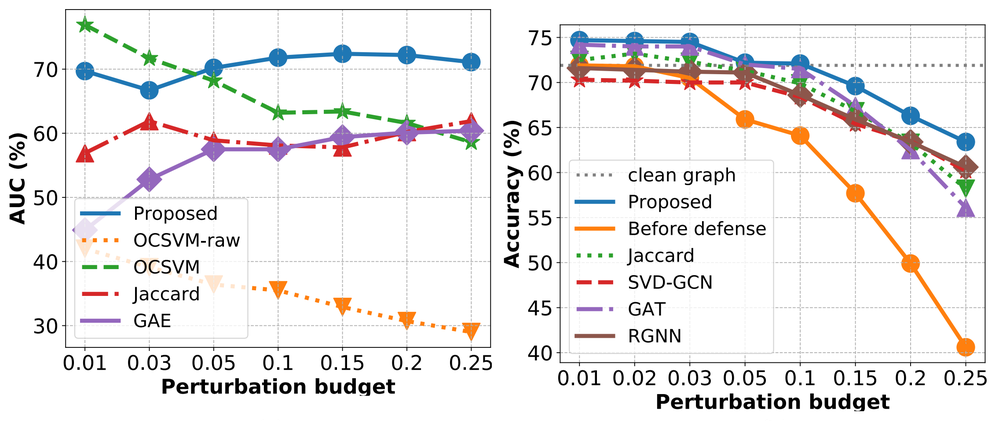
Dataset	Ours	GraphSAC	GAE	Amen	Radar	OCSVM raw	OCSVM emb	Jaccard
Cora	**86.4**	80.0	50.2	75.0	77.0	50.3	72.7	69.9
Citeseer	**80.1**	75.0	64.4	73.0	67.0	36.8	67.8	69.3
Polblogs	85.4	**98.0**	51.2	89.0	76.0	-	59.9	-
Pubmed	**87.8**	82.0	69.2	62.0	44.0	58.5	57.9	82.4

- **Adaptive graph adversarial attacks**
- Goal: Examine under the assumption that the attacker aware our detector's existence, how well the detector performs.

	Clean Nettack	Nettack	Nettack mJSD	Clean Dice	DICE	DICE mJSD
Cora	82.2	25.0	49.3	89.8	64.1	63.2
Citeseer	73.6	28.6	45.0	87.5	62.6	61.8
Polblogs	95.0	37.8	84.3	87.0	27.7	64.8
Pubmed	89.2	9.7	52.4	86.7	58.4	61.1

Accuracy (%) comparison for the target nodes between the original attacks and the adaptive attacks with the multi-JSD constraint.

- **Defense strategy based on local signal smoothness**
  - Goal: Examine the effectiveness of the proposed defense solution [deactivate the neighbor information for the flagged nodes]



(a) (b)

Detection and defense under different perturbation budget (Citeseer)

- **Real world application: Adversarial attack detection in Bitcoin Networks**

(Acc %)	Clean	Under Nettack	Defense Nettack	Detection Nettack (AUC %)
TS 1	92.5	85	**87.5**	83.5
TS 2	72.5	50.0	**52.5**	78.0
TS 3	90.0	62.5	**82.5**	86.3
TS 4	97.5	67.5	**82.5**	87.1
TS 5	95	55.0	**87.5**	75.9

Left: Prediction accuracy (%) for the illicit transaction before and after the adversarial graph perturbations and after using our defense strategy. Right: Detection Area-under-Curve (AUC) under Nettack attacks using GCN as the prediction model.

Conclusion

- We have presented methods for detecting adversarial attacks against graph data.
- Building on the detection procedures, we designed a defense mechanism that results in a much more robust training procedure.
- We proposed an adaptive attack which significantly reduces the detection rate by using our designed smoothness metric as an unnoticeable criteria for limiting the search space for the perturbation edge.

References

[1] Kipf, Thomas and Welling, Max. 2017. Semi-supervised classification with graph convolutional networks. In *Proc. ICLR*.
[2] Veličković, Petar et al. 2018. Graph Attention Networks. In *Proc. ICLR*.
[3] Zhang, Yingxue et al. 2019. Bayesian Graph Convolutional Neural Networks for Semi-supervised Classification. In *Proc. AAAI*.
[4] Zügner et, Daniel al. Adversarial Attacks on Neural Networks for Graph Data. In *Proc. KDD*.
[5] Zügner, Daniel and Günnemann, Stephan. 2019. Adversarial Attacks on Graph Neural Networks via Meta Learning. In *Proc. ICLR*.