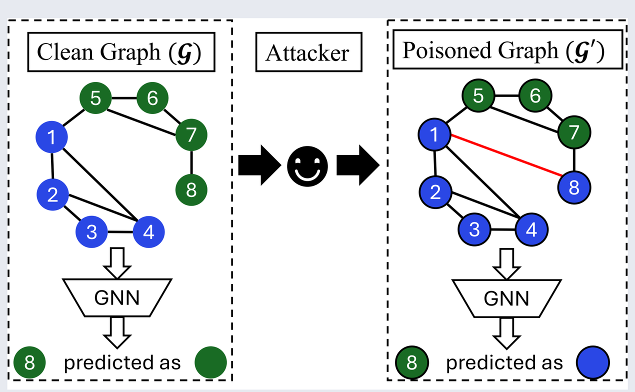
When Witnesses Defend: A Witness Graph Topological Layer for Adversarial Graph Learning

Naheed Anjum Arafat*, Debabrota Basu+, Yulia R. Gel#, Yuzhou Chen$

*Nanyang Technological Univ. (Singapore), +Univ. Lille, Inria (France), #Virginia Tech (USA), $ UC Riverside (USA)

Adversarial Attack on GNNs

The attacker **misleads GNNs** into making incorrect predictions by deliberately perturbing **a small number of edges** (e.g. remove/add edges) or node features.



Goal: Design a robust representation R such that

$$\|R(\mathcal{G}) - R(\mathcal{G}')\|_1 = \mathcal{O}(\delta)$$
$$\text{when } \|\mathcal{G} - \mathcal{G}'\|_1 = \delta$$

Our Contributions

- This is the first work that shows that **topological features can make GNNs robust** against adversarial attacks.
- Our approach integrates **local** and **global** higher-order graph characteristics (**using Persistence homology**) and controls their potential defense role via a **topological regularizer**.
- Effective against a wide **variety of attacks**:
  - Targetted (perturbs neighbors of a set of target nodes)
  - Global (perturbs whichever edges minimizes the attackers loss)
  - Adaptive (White-box, the model architecture, parameters and the defense mechanisms are known to the attacker)
  - Node feature attack

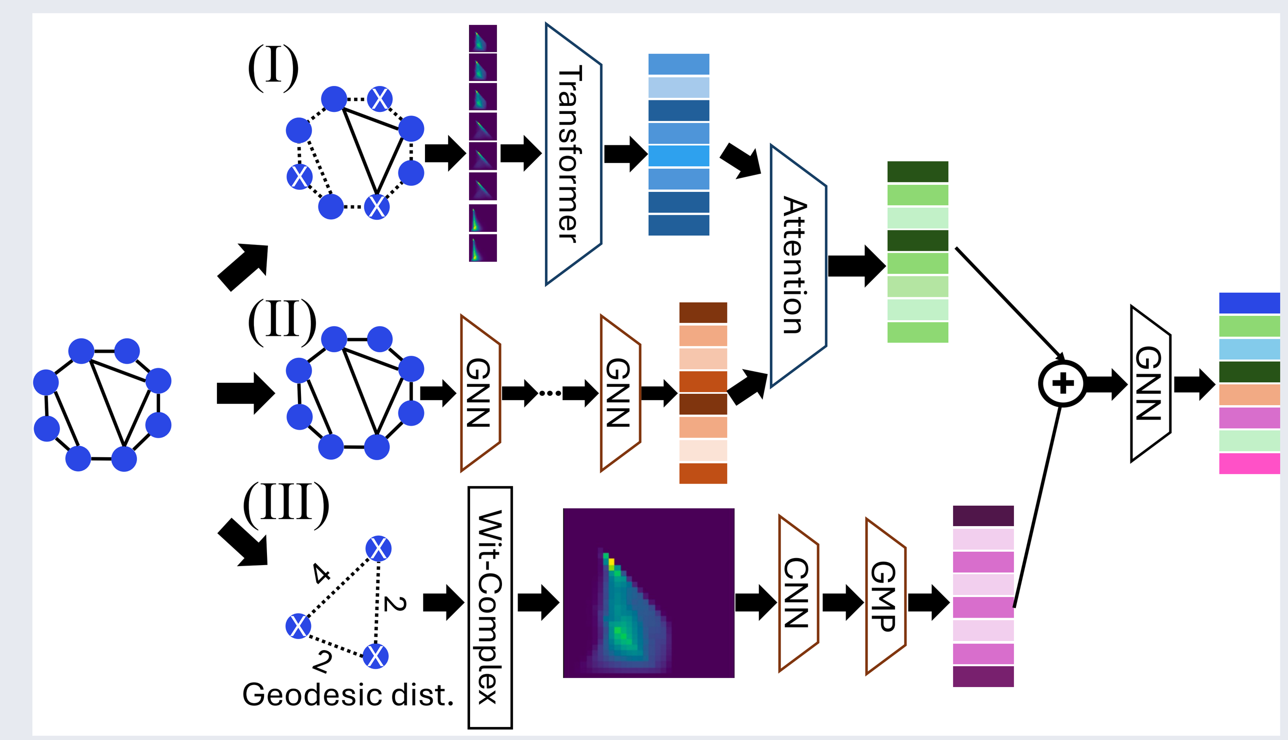
Performance (Mettack)

Dataset	Models	Perturbation Rate 0%	Perturbation Rate 10%
Cora-ML	GCN	82.87±0.83	70.39±1.28
	GCN+WGTL$_P$(ours)	**83.83±0.55**	**71.31±0.85**
	GAT	84.25±0.67	72.63±1.56
	GAT+WGTL$_P$(ours)	****86.07±2.10**	****73.74±1.92**
	GraphSAGE	81.00±0.27	70.92±1.18
	GraphSAGE+WGTL$_P$(ours)	**83.63±0.35**	**73.57±0.73**
	ProGNN	82.98±0.23	71.59±1.33
	ProGNN+WGTL$_P$(ours)	**83.85±0.38**	**72.71±1.26**
	GCN+GNNGuard	83.21±0.34	69.13±0.77
	GCN+GNNGuard+WGTL$_P$(ours)	**84.78±0.43**	**70.15±0.89**
Polblogs	GCN	94.40±1.47	69.16±1.86
	GCN+WGTL$_P$(ours)	**95.95±0.15**	**74.52±0.28**
	GAT	95.28±0.51	73.11±1.20
	GAT+WGTL$_P$(ours)	**95.87±0.26**	**74.21±0.74**
	GraphSAGE	94.54±0.27	74.66±0.85
	GraphSAGE+WGTL$_P$(ours)	**95.58±0.50**	****74.93±0.81**
	GCN+GNNGuard	95.03±0.25	72.76±0.75
	GCN+GNNGuard+WGTL$_P$(ours)	****96.22±0.25**	**73.72±1.00**

Efficiency comparison

Datasets/ (# Landmarks)	Landmark selection time (s)	Local feat. comput. time (s)	Global feat. comput. time (s)
Cora-ML/124	0.01±0.01	0.12±0.03	5.11±0.13
Citeseer/105	0.01±0.01	0.16±0.02	5.23±1.22
Polblogs/61	0.01±0.00	0.07±0.01	4.64±0.2
Snap-patents/91	0.03±0.02	0.64±0.00	7.54±1.15
Pubmed/394	0.07±0.01	0.51±0.03	27.83±0.47
OGBN-arXiv/84	1.02 ±0.00	12.79±0.31	83.04±2.19

Witness Graph Topological Layer (WGTL)



- **Ⅰ Local Topology Encoding:** Encodes local topological features of every node. ($\boldsymbol{Z}_{T_L}$)
- **Ⅱ Node Representation Learning**. Learns node representations using any backbone GNN. ($\boldsymbol{Z}_G$)
- **Ⅲ Global Topology Encoding.** Encodes topological features of the entire graph. ($\boldsymbol{Z}_{T_G}$)
- **Ⅳ Aggregated Topological Encoding.** Encodes local and global topological priors. ($\boldsymbol{Z}_{\text{WGTL}}$)

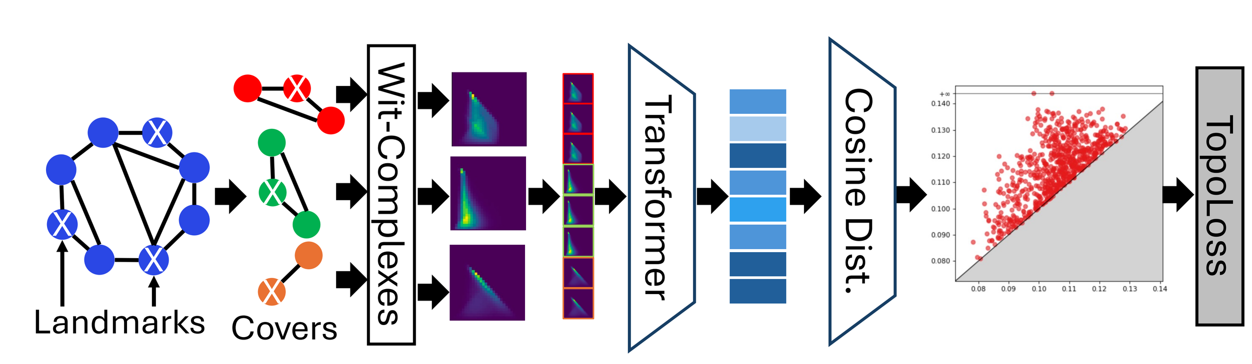
$$z = [\boldsymbol{Z}_{T_L}, \boldsymbol{Z}_G]$$

$$\text{Attention coefficients, } \alpha_i = \textbf{Softmax}(\boldsymbol{W}_2 \cdot \tanh(\boldsymbol{W}_1 z_i + \boldsymbol{b}_1))$$

$$\text{Additive attention, } \boldsymbol{Z}_{\text{AGG}} = \alpha_1 \times \boldsymbol{Z}_{T_L} + \alpha_2 \times \boldsymbol{Z}_G$$

$$\boldsymbol{Z}_{\text{WGTL}} = \boldsymbol{Z}_{\text{AGG}} \boldsymbol{Z}_{T_G}$$

Topological Regularizer



$$L_{topo}(\mathrm{T}(\mathcal{G})) \triangleq \sum_{i=1}^{m} (d_i - b_i)^2 \left(\frac{d_i + b_i}{2}\right)^2$$

- A localized attack appears as topological noise in the final persistent diagram (PD), and exhibit lower persistence.
- And minimising L_{topo} forces the Transformer to learn local topology encodings (Z_{T_L}) which produces PD with small persistence, i.e., $(d_i - b_i)$.

Conclusion

- Proposed the first topological defense against adversarial attacks on graphs.
- We have derived theoretical properties of WGTL, both at the local and global levels.
- WGTL improves robustness against a wide variety of attacks.

Acknowledgements: NSF grant TIP-2333703, ONR grant N00014-21-1-2530, ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01), PEPR project FOUNDRY (ANR23-PEIA-0003), and the CHIST-ERA project CausalXRL (ANR-21-CHR4-0007).

arXiv Link