



TopFeaRe: Locating Critical State of Adversarial Resilience for Graphs Regarding Topology-Feature Entanglement

Xinxin Fan[1,2], Wenxiong Chen[3,1], Quanliang Jing[1], Chi Lin[3], Shaoye Luo[1,2], Wenbo Song[3,1], Yunfeng Lu[4]

[1] State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences

[2] University of Chinese Academy of Sciences

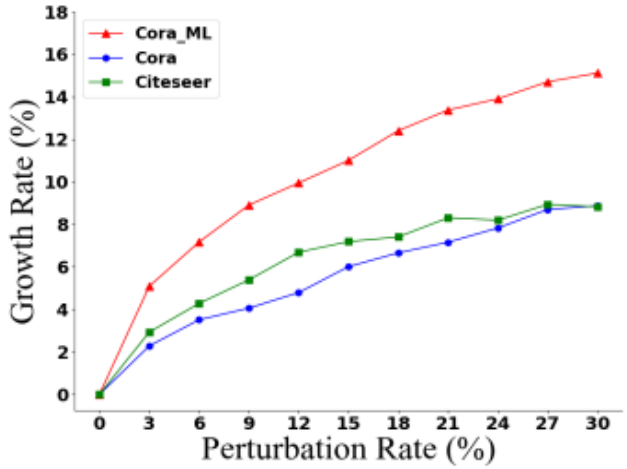
[3] Dalian University of Technology

[4] Beihang University

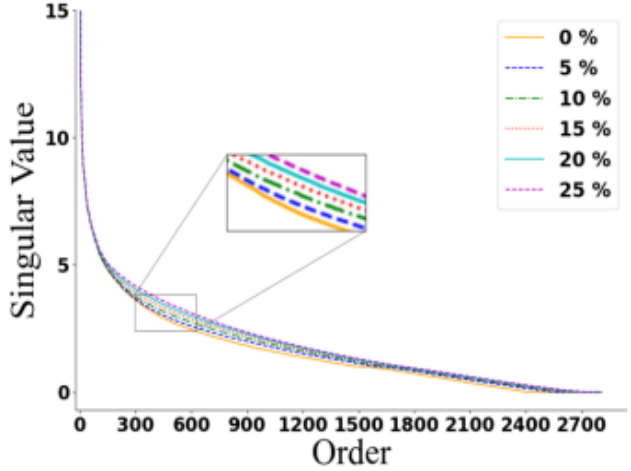
Outline

- Background
- Questions
- Methodology
- Experiments
- Conclusion

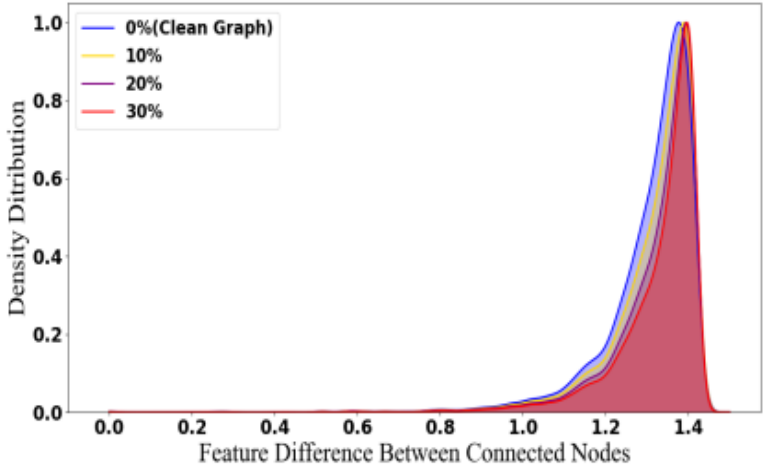
Background



(a) Rank Growth



(b) Singular Value



(c) Feature Smoothness

Variations of adjacency-matrix rank and singular value, feature smoothness under attack Metattack

Observations:

i) Rank of adjacency matrix goes up as the rate of adversarial perturbation enlarges on Cora/Cora_ML/Citeseer

ii) Singular value of adjacency matrix raises, especially in the middle segments

iii) Density distribution of feature difference over connected nodes has a right-transfer after attacks

Questions

RQ-I: Why the existing preprocessing-based methods are effective and whether the original characteristics might be mistakenly deleted

RQ-II: Whether to achieve the true meaning of jointly-learning of graph topology and node features

RQ-III: What the relationship between graph data and neural-network learning is

Methodology

Whether there exists a critical attack-resilient state to resist adversarial attacks?

To this end, regarding topology and feature together, we employ equilibrium-point theory to deduce the asymptotically-stable equilibrium point by which the necessarily-removed edges are inferred to boost the attack resilience of perturbed graphs.

Methodology

☐ Adversarial Attack Modeling

From the standpoint of adversarial perturbations, we provide a universal differential-form definition to sketch the dynamics under graph adversarial perturbations:

$$\frac{d}{dt}\vec{x}(t) = \mathbf{A}\vec{x}(t) + \mathbf{B}\left(-\vec{\phi}\left(\vec{y}(t)\right)\right)$$
$$\vec{y}(t) = \mathbf{C}\vec{x}(t)$$

$\frac{d}{dt}\vec{x}(t)$ indicates the state variance of all nodes during the adversarial perturbations over time *t*.
$A\vec{x}(t)$ denotes *N*-dimensional vector denoting the affected linear perturbations.
$\vec{y}(t)$ is the output vector standing for the states of all nodes afterward the one-round adversarial perturbations.
Matrices ***A***, ***B***, ***C*** are coefficients demonstrating detailed dynamics of graph during adversarial learning.

Methodology

☐ Adversarial Attack Modeling

The node-state dynamic-variation under adversarial attacks can be deduced as:

$$\frac{dr_{Gra}}{dt} = -\mathcal{M} r_{Gra} + \frac{\langle \gamma_r \rangle (q_{Gra})^h}{1 + \langle \gamma_r \rangle (q_{Gra})^h}$$
$$\frac{dq_{Gra}}{dt} = -\mathcal{C} q_{Gra} + \frac{\langle \gamma_q \rangle (r_{Gra})^h}{1 + \langle \gamma_q \rangle (r_{Gra})^h} .$$

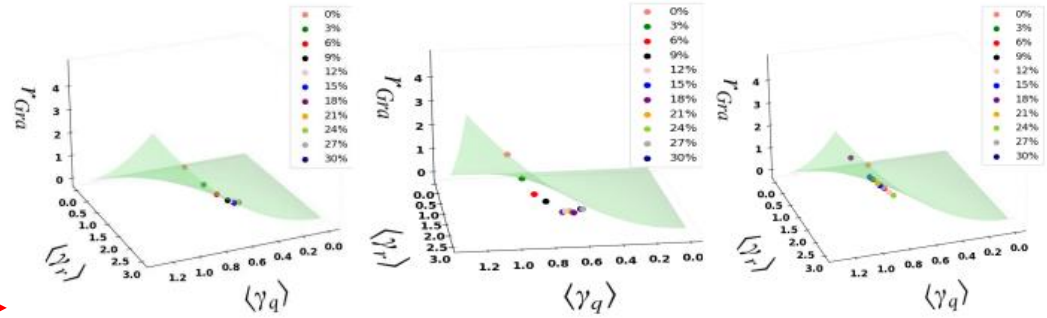
$\mathcal{M}$ -- average degree centrality.
$\mathcal{C}$ -- average feature distinction.
r_{Gra} -- graph-topology dynamics.
q_{Gra} -- node-feature dynamics.
h -- adaptive parameters representing the node-state dynamics.

Methodology

☐ Critical Attack-Resilient State

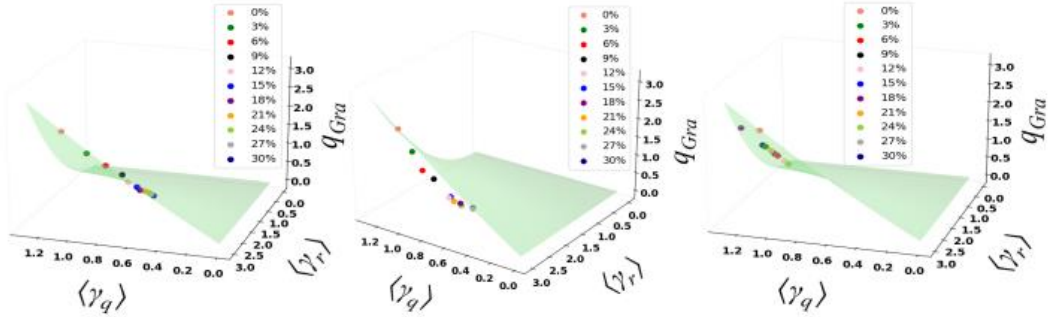
Asymptotically-Stable Equilibrium Point (ASEP) :
From **Theory** to **Reality**

Theorem 4.3. ***(Asymptotic Stability of Two-Dimensional Dynamic-Variation Mapping Equation)*** *Decompose the two-dimensional mapping function (Eq. (19)) into two the one-dimensional and analyze the asymptotic stability of each separately. According to the structure of Eq. (1), let* $\mathbf{A}_r$ *and* $\mathbf{B}_r$ *correspond to* $-\mathcal{M}$ *and* $\langle \gamma_q \rangle$ *in the dynamic equations describing the state of graph topology, while* $\mathbf{A}_q$ *and* $\mathbf{B}_q$ *correspond to* $-\mathcal{C}$ *and* $\langle \gamma_r \rangle$ *in the dynamic equations describing the state of node feature. As long as* $\mathbf{A}_r$, $\mathbf{B}_r$, $\mathbf{A}_q$ *and* $\mathbf{B}_q$ *can form strictly positive real matrix* $\frac{1}{k_r} + (I + s\psi)\,\mathbf{C}\,(sI - \mathbf{A}_r)^{-1}\,\mathbf{B}_r$ *and* $\frac{1}{k_q} + (I + s\psi)\,\mathbf{C}\,(sI - \mathbf{A}_r)^{-1}\,\mathbf{B}_r$, *the two-dimensional dynamic equation exhibits asymptotic stability.*



(a) Metattack (b) CE-PGD (c) DICE

Figure 2: ASEP surface with r_{Gra} on Cora_ML



(a) Metattack (b) CE-PGD (c) DICE

Figure 3: ASEP surface with q_{Gra} on Cora_ML

Experiments

Table 2: Accuracy of node classification on Cora_ML (**Bold**-the best)

GAA	Mod. \ RAP	0%	5%	10%	15%	20%	25%
Metattack	GCN-SVD	0.8271±0.0080	0.8164±0.0084	0.8054±0.0092	0.7857±0.0095	0.6967±0.1935	0.6279±0.1729
	GCN-Jaccard	0.8423±0.0094	0.7896±0.0087	0.7441±0.0123	0.7028±0.0145	0.6552±0.0294	0.5923±0.0471
	TopFeaRe	**0.8524±0.0096** (↑ **1.01%**)	**0.8293±0.0088** (↑ **1.29%**)	**0.8219±0.0114** (↑ **1.65%**)	**0.8150±0.0107** (↑ **2.93%**)	**0.8086±0.0104** (↑ **11.19%**)	**0.7920±0.0128** (↑ **16.41%**)
CE-PGD	GCN-SVD	0.8271±0.0080	0.8229±0.0066	0.8164±0.0080	0.8118±0.0122	0.8110±0.0089	0.8080±0.0055
	GCN-Jaccard	0.8423±0.0094	0.8334±0.0106	0.8248±0.0112	0.8153±0.0086	0.8131±0.0109	0.8117±0.0082
	TopFeaRe	**0.8524±0.0096** (↑ **1.01%**)	**0.8357±0.0092** (↑ **0.23%**)	**0.8327±0.0112** (↑ **0.79%**)	**0.8307±0.0086** (↑ **1.54%**)	**0.8305±0.0092** (↑ **1.74%**)	**0.8294±0.0090** (↑ **1.77%**)
DICE	GCN-SVD	0.8271±0.0080	0.8148±0.0065	0.8000±0.0064	0.7901±0.0089	0.7730±0.0119	0.7590±0.0087
	GCN-Jaccard	0.8423±0.0094	0.8300±0.0092	0.8177±0.0101	0.8061±0.0088	0.7915±0.0135	0.7752±0.0128
	TopFeaRe	**0.8524±0.0096** (↑ **1.01%**)	**0.8354±0.0089** (↑ **0.54%**)	**0.8291±0.0113** (↑ **1.14%**)	**0.8242±0.0103** (↑ **1.81%**)	**0.8160±0.0122** (↑ **2.45%**)	**0.8078±0.0079** (↑ **3.26%**)

Table 3: Accuracy of node classification on Citeseer under combination with Ours using Metattack

Data set	Mod. \ RAP	0%	5%	10%	15%	20%	25%
Citeseer	GCN	0.7211±0.0099	0.6661±0.0219	0.6132±0.0325	0.5538±0.0273	0.4935±0.0261	0.4556±0.0318
	TopFeaRe	**0.7228±0.0105** (↑ **0.17%**)	**0.7159±0.0170** (↑ **4.98%**)	**0.7005±0.0162** (↑ **8.73%**)	**0.6885±0.0170** (↑ **13.47%**)	**0.6754±0.0147** (↑ **18.19%**)	**0.6647±0.0189** (↑ **20.91%**)
	GAT	0.7335±0.0084	0.7140±0.0161	0.6876±0.0256	0.6511±0.0301	0.6170±0.0203	0.5960±0.0283
	GAT-Our	**0.7393±0.0127** (↑ **0.58%**)	**0.7274±0.0094** (↑ **1.34%**)	**0.7210±0.0120** (↑ **3.34%**)	**0.7105±0.0099** (↑ **5.94%**)	**0.6991±0.0137** (↑ **8.21%**)	**0.6934±0.0201** (↑ **9.74%**)
	HANG	**0.7406±0.0116**	0.7268±0.0117	0.7075±0.0272	0.6851±0.0156	0.6722±0.0214	0.6678±0.0305
	HANG-Our	0.7399±0.0104 (↓ **0.07%**)	**0.7348±0.0126** (↑ **0.80%**)	**0.7281±0.0118** (↑ **2.06%**)	**0.7160±0.0106** (↑ **3.09%**)	**0.7095±0.0145** (↑ **3.73%**)	**0.7093±0.0110** (↑ **4.15%**)
	MidGCN	0.7452±0.0032	0.7237±0.0067	0.7105±0.0066	0.6584±0.0106	0.6700±0.0052	0.6306±0.0099
	MidGCN-Our	**0.7466±0.0047** (↑ **0.14%**)	**0.7307±0.0027** (↑ **0.70%**)	**0.7127±0.0080** (↑ **0.22%**)	**0.6626±0.0102** (↑ **0.42%**)	**0.6712±0.0054** (↑ **0.12%**)	**0.6349±0.0134** (↑ **0.43%**)

Comparison with Preprocessing-based Baselines:

--On Cora_ML under RAP of 25%, TopFeaRe outperforms the best baseline by 16.41%, 1.77%, and 3.26%

-- At RAP 0% (no attack), TopFeaRe can further promote the adversarial resilience of the clean graph

TopFeaRe not only performs exceptionally well in adversarial environments, but it also enhances the learning capability of the graph topology under clean conditions.

Comparison with Neural Network-Optimized Baselines:

RAP ranging from 5%-25%, with the average accuracy promotion of GCN, GAT, HANG, and Mid-GCN

-- [13.095%, 8.67%, 1.06%, 0.43%] on Cora_ML

-- [18.75%, 7.78%, 4.04%, 1.81%] on Cora

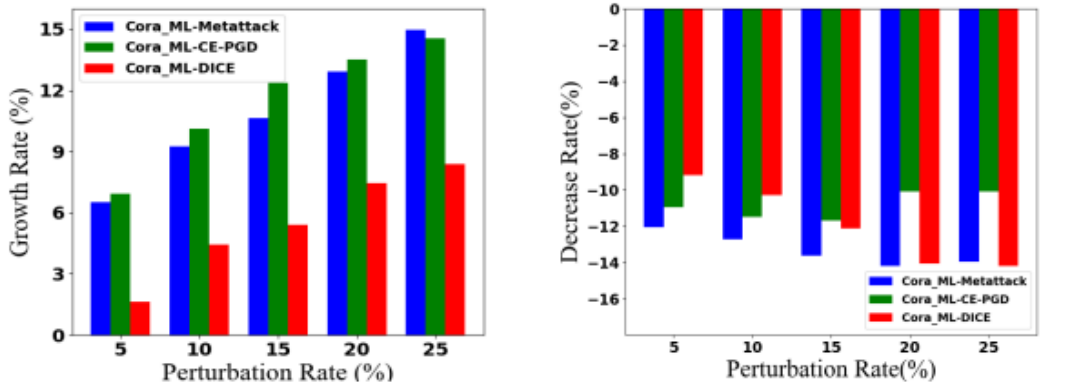
-- [13.26%, 5.71%, 2.77%, 0.38%] on Citeseer

-- [4.54%, 9.39%, 1.01%, 1.23%] on Amazon Photo

-- [7.30%, 14.86%, 7.29%, 2.26%] on PubMed

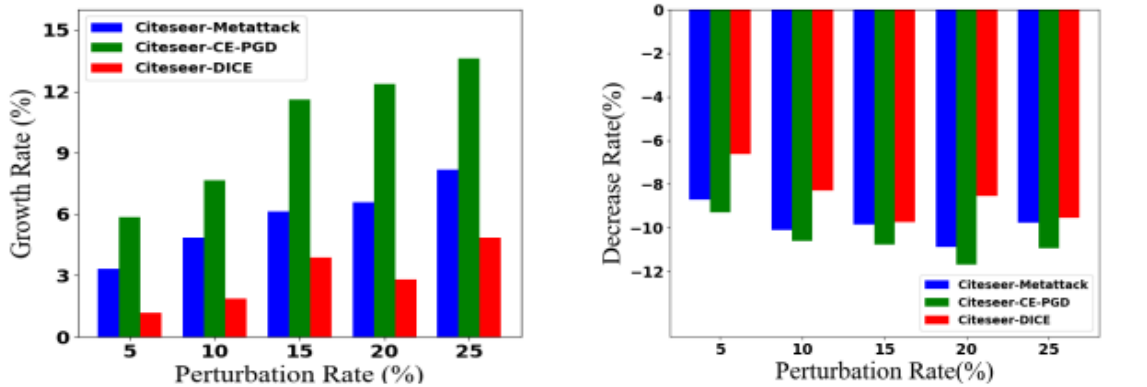
This indicates TopFeaRe can significantly boost the adversarial resilience of the neural network-optimized baselines.

Experiments



(a) Rank growth under attacks (b) Rank decreased by TopFeaRe

Figure 4: Rank variation of adjacency matrix of Cora_ML.



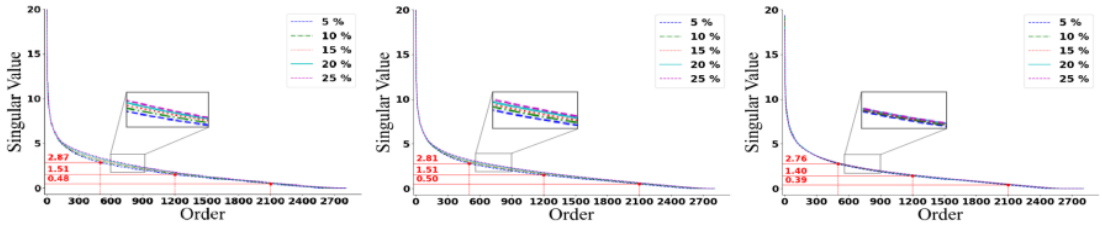
(a) Rank growth under attacks (b) Rank decreased by TopFeaRe

Figure 5: Rank variation of adjacency matrix of citeseer.

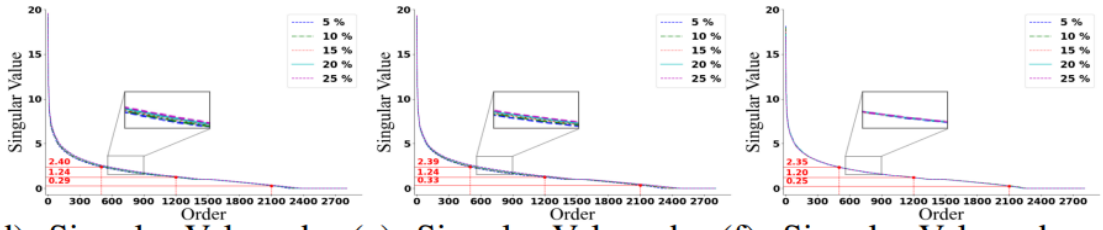
As RAP increases from 5% to 25%, the rank of adjacency matrix correspondingly raises.
This phenomenon suggests:
-- Adversarial attacks tend to establish connections between nodes, leading to a more complex and dense graph structure
-- In response, TopFeaRe significantly reduces the rank through effectively removing the edges over dissimilar nodes

This process not only mitigates the impact of adversarial perturbations, but also optimizes graph's topology, making it closer to a robust state.

Experiments

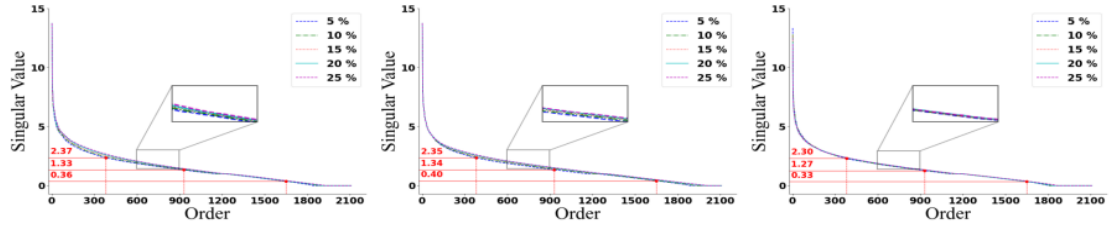


(a) Singular-Value under Metattack (b) Singular-Value under CE-PGD (c) Singular-Value under DICE

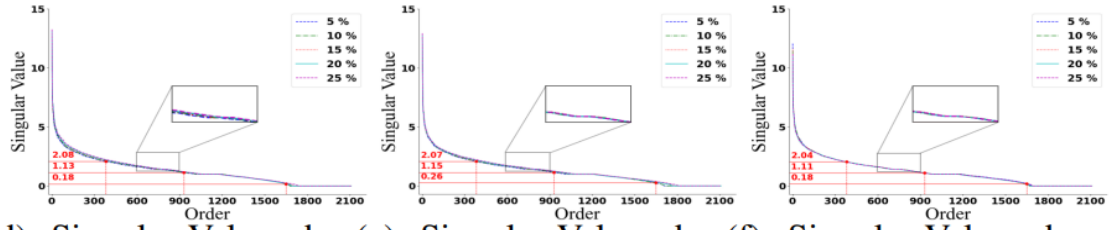


(d) Singular-Value decrease (w.r.t. Metattack) (e) Singular-Value decrease (w.r.t. CE-PGD) (f) Singular-Value decrease (w.r.t. DICE)

Figure 6: Singular-Value variation of Cora_ML



(a) Singular-Value under Metattack (b) Singular-Value under CE-PGD (c) Singular-Value under DICE



(d) Singular-Value decrease (w.r.t. Metattack) (e) Singular-Value decrease (w.r.t. CE-PGD) (f) Singular-Value decrease (w.r.t. DICE)

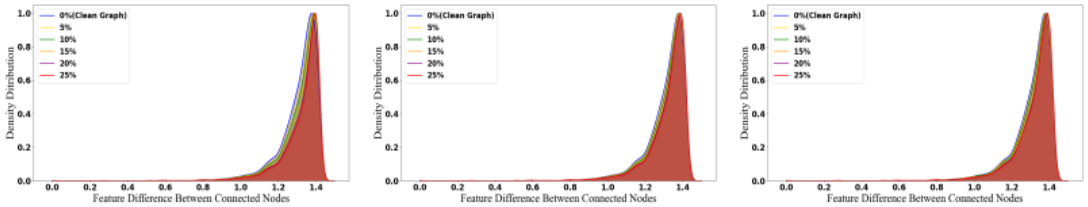
Figure 7: Singular-Value variation of Citeseer

Singular values are part of the spectral properties of a graph, reflecting its structural information.

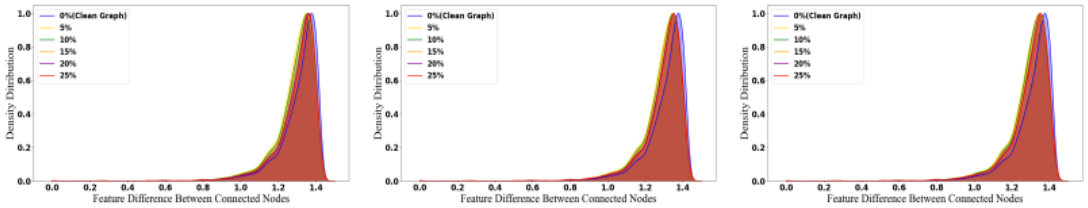
-- GAAs can alter the spectral properties of a graph, making an originally simple graph more complex to deceive target models, which usually renders the increase in singular values

-- TopFeaRe successfully reduces the singular values for both datasets, effectively mitigating the impact of adversarial attacks

Experiments

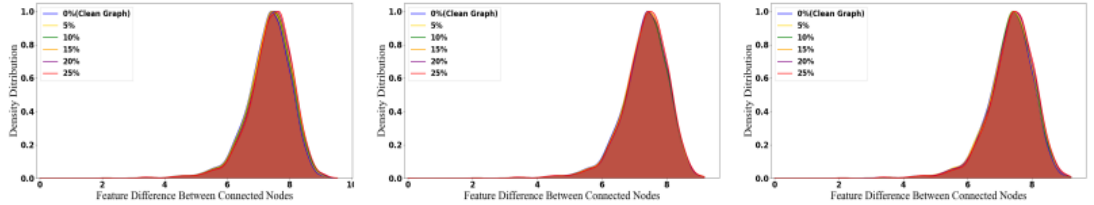


(a) Feature-Smoothness under Metattack (b) Feature-Smoothness under CE-PGD (c) Feature-Smoothness under DICE

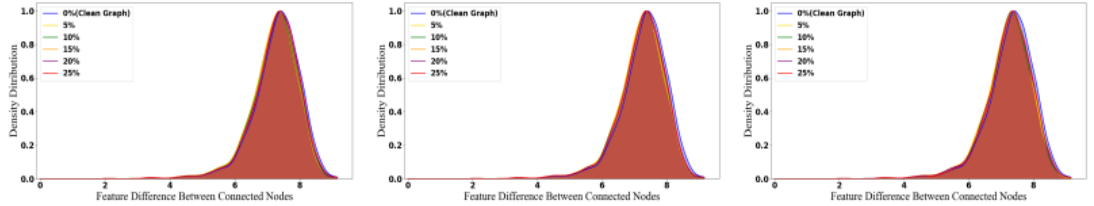


(d) Feature-Smoothness: Metattack vs. TopFeaRe (e) Feature-Smoothness: CE-PGD vs. TopFeaRe (f) Feature-Smoothness: DICE vs. TopFeaRe

Figure 8: Feature-Smoothness variation of Cora_ML



(a) Feature-Smoothness under Metattack (b) Feature-Smoothness under CE-PGD (c) Feature-Smoothness under DICE



(d) Feature-Smoothness: Metattack vs. TopFeaRe (e) Feature-Smoothness: CE-PGD vs. TopFeaRe (f) Feature-Smoothness: DICE vs. TopFeaRe

Figure 9: Feature-Smoothness variation on Citeseer

Adversarial attacks causing feature non-smoothness can mislead the target models during both training and inference.

-- In clean graphs feature values are typically concentrated in regions with low feature distinctions, whereas after perturbations, it is reversed

-- TopFeaRe causes the edges in high feature-distinction regions to disperse, while features become concentrated in areas with low distinctions

Conclusion

- Exhibit what aspects are influenced by GAA, and theoretically explain why target models are easily deceived by adversarial perturbations
- Design topology-feature-entangled function to model graph's dynamics under adversarial attacks
- Formally analyze what prerequisites are required to satisfy the existence theory of asymptotically-stable equilibrium point

Thank you for your attention!

Q/A?



