On Which Nodes Does GCN Fail? Enhancing GCN From the Node Perspective

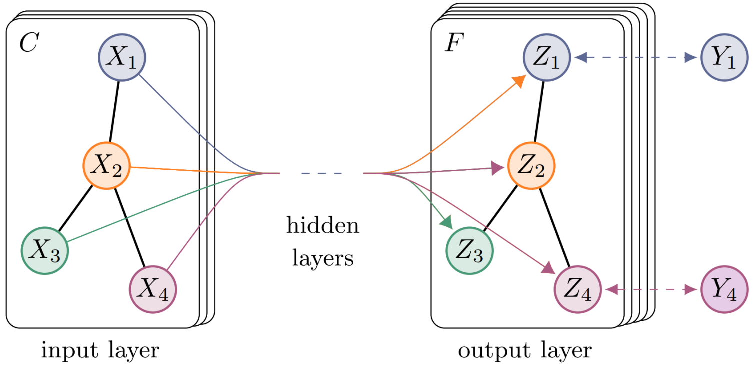
Jincheng Huang[1], Jialie Shen[2], Xiaoshuang Shi[1]*, Xiaofeng Zhu[1]*

1.University of Electronic Science and Technology of China 2. City, University of London, London

Background

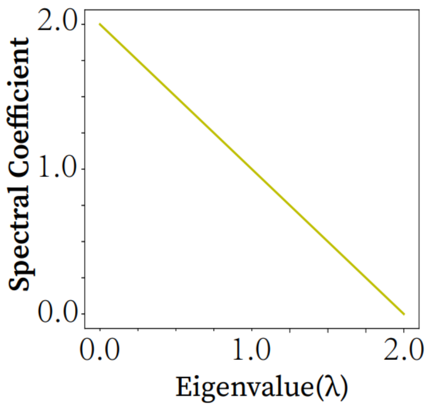
Model

Graph Convolutional Networks (GCNs)



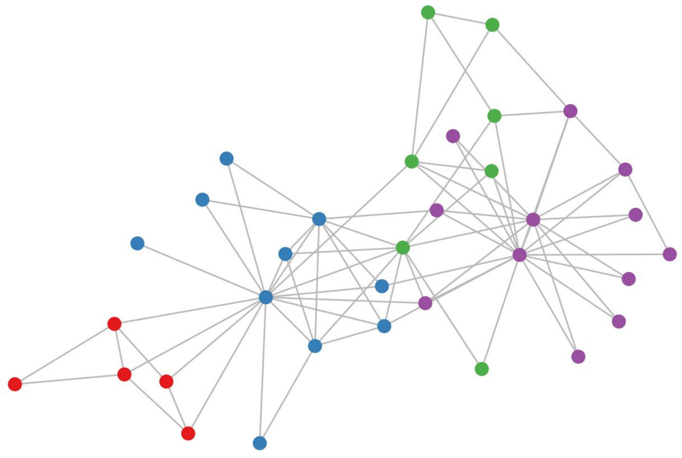
Operation

Feature Smoothing



Result

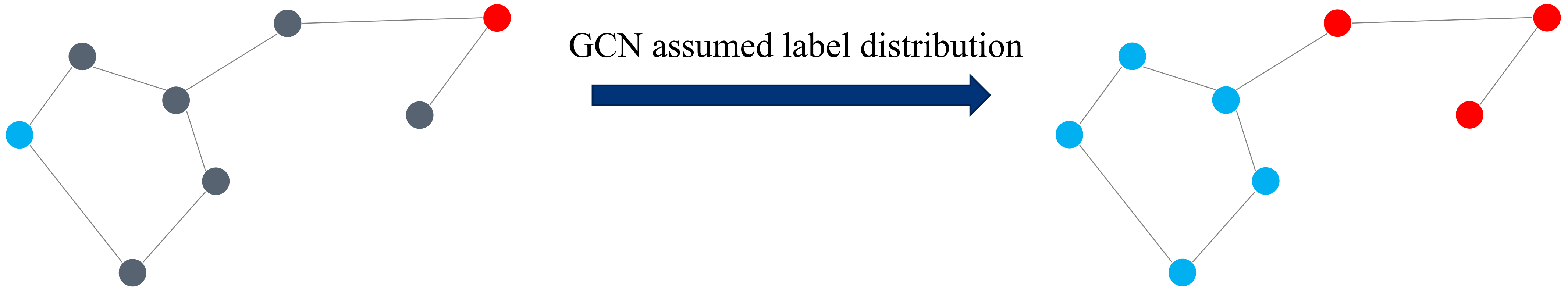
Connected nodes have similar features



GCNs excel at handling graph-structured data, with most methods relying on their feature smoothing operations.

Background

What kind of graph data does GCN expect?



GCNs assume that Connect nodes are highly likely to share the same labels. (i.e., label smoothness assumption)(Zhang et al., 2021)

Question: Is the label distribution obtained by GCN feature smoothing consistent with the label smoothness assumption?

Background

University of Electronic Science and Technology of China

Theorem 1

For nodes with unknown labels in the graph, the upper bound of the GCN's generalization ability reaches optimal if the true labels of these nodes are equal to the labels generated by the LPA.

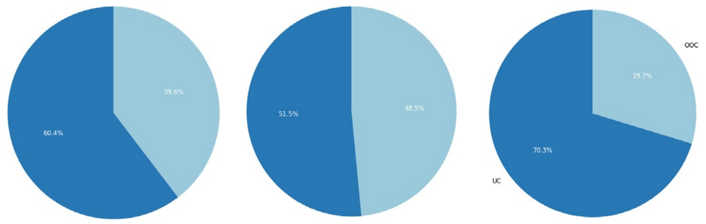
- Theorem 1 establishes the link between the output of LPA and the expected label distribution of GCN (i.e., label smoothness assumption)

Label-Feature Smoothing Alignment Algorithm

1. GCN feature smoothing: $\mathbf{Y}_{fs} = \widehat{\mathbf{A}}^L MLP(\mathbf{X})$ 2. GCN label smoothness assumption: $\mathbf{Y}_{lp} = \widehat{\mathbf{A}}^L \mathbf{Y}$

$\mathbf{V}_{OOC} = \{\mathbf{V}_i | argmax(\mathbf{Y}_{fs,i}) \neq argmax(\mathbf{Y}_{lp,i}), i \in [n]\}$ $\mathbf{V}_{UC} = \mathbf{V} - \mathbf{V}_{OOC}$

Answer the Question: There is a fairly significant proportion of nodes that are not consistent.



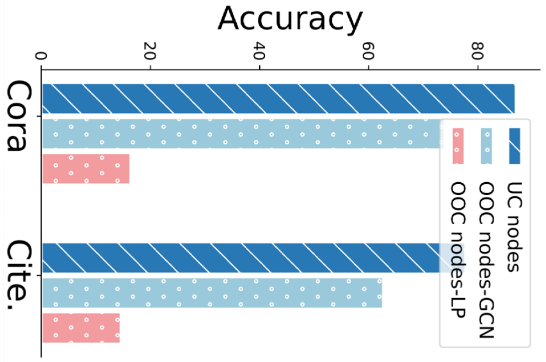
Cora Citeseer Pubmed

Background

Unlabeled Nodes

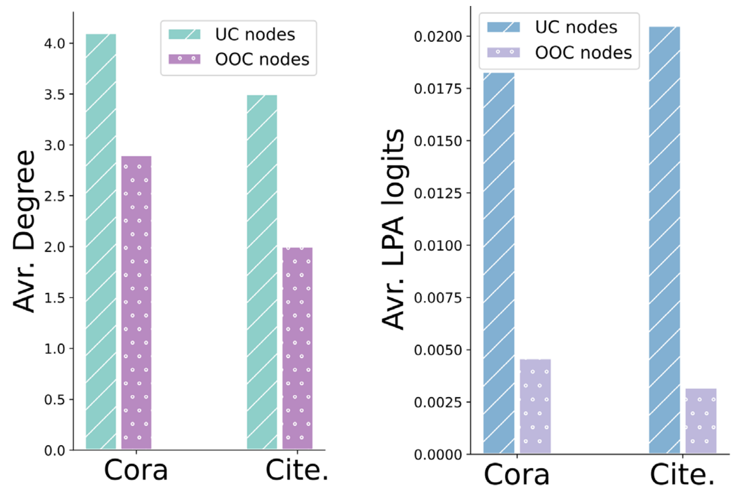
UC Nodes

Nodes that achieve label smoothing assumptions using GCN feature smoothing operations are under the control of GCN.



OOC Nodes

Nodes affected by GCN's feature smoothing operation conflict with the label smoothness assumption, making it difficult to correct representation under the GCN framework.



Character of OOC nodes.

(i) Nodes with few neighbors (left figure).

(ii) Nodes away from labeled nodes (right figure).

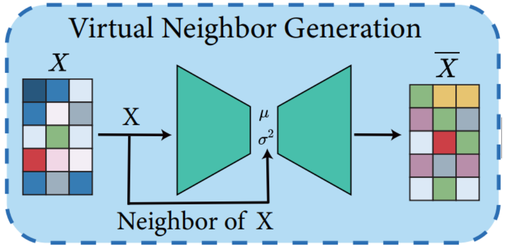
Method

For Nodes with Few Neighbors

Virtual Neighbor Generation

Use $X_v (v \in V)$ as a condition, and to learn the neighbor distribution of $X_u (u \in N_v)$ (Liu et al., 2022, Sohn et al., 2015).

$$\mathcal{L}_{ELBO} = -KL(q(\mathbf{z} \mid \mathbf{X}_u \mathbf{X}_v) \| p(\mathbf{z} \mid \mathbf{X}_v)) + \mathbb{E}_{q(\mathbf{z}|\mathbf{X}_u,\mathbf{X}_v)}(p(\mathbf{X}_u \mid \mathbf{X}_v, \mathbf{z}))$$

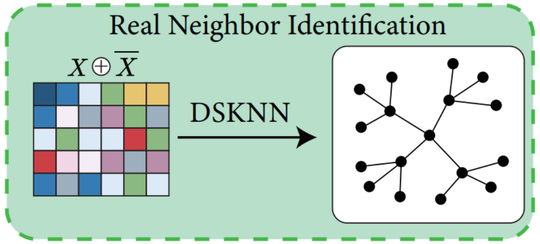


This process allows us to obtain the node v's virtual neighbor feature vector $\overline{\mathbf{X}}_v$.

Potential Real Neighbor Identification

Virtual nodes contain only first-order information and can't affect message passing. We posit potential non-directly connected neighbors can augment message passing for OOC nodes if:

- They are in the same subspace.
- Their neighbors are in the same subspace.



$$\min_{\mathbf{S}} \sum_{i,j=0}^{n} \left(-s_{i,j} \mathbf{X}_i^T \mathbf{X}_j - s_{i,j} \overline{\mathbf{X}}_i^T \overline{\mathbf{X}}_j + s_{i,j}^2 \right)$$

$$s_{i,j} = \frac{1}{2} \left(\mathbf{X}_i^T \mathbf{X}_j + \overline{\mathbf{X}}_i^T \overline{\mathbf{X}}_j \right) = \frac{1}{2} \left(\mathbf{X}_i \oplus \overline{\mathbf{X}}_i \right)^T \left(\mathbf{X}_j \oplus \overline{\mathbf{X}}_j \right)$$

Method

Nodes away from labeled nodes

Theorem 2

Given an undirected graph G(V, E) has n nodes and m edges. Assuming there are q nodes in the graph with labels selected uniformly at random. The occurrence probability of nodes that are not affected by labels with a two-layer GCN is equal to

$$(1-\frac{q}{n})(1-\frac{q}{n-1})\prod_{i=1}^{q}(1-\frac{2m}{n(n-1)-2i})\prod_{i=q}^{2q}(1-\frac{2(m-1)}{n(n-1)-2i})$$

- Theorem 2 tells us the occurrence probability of unaffected by labeled nodes is negatively correlated with the number of labels and total edges.

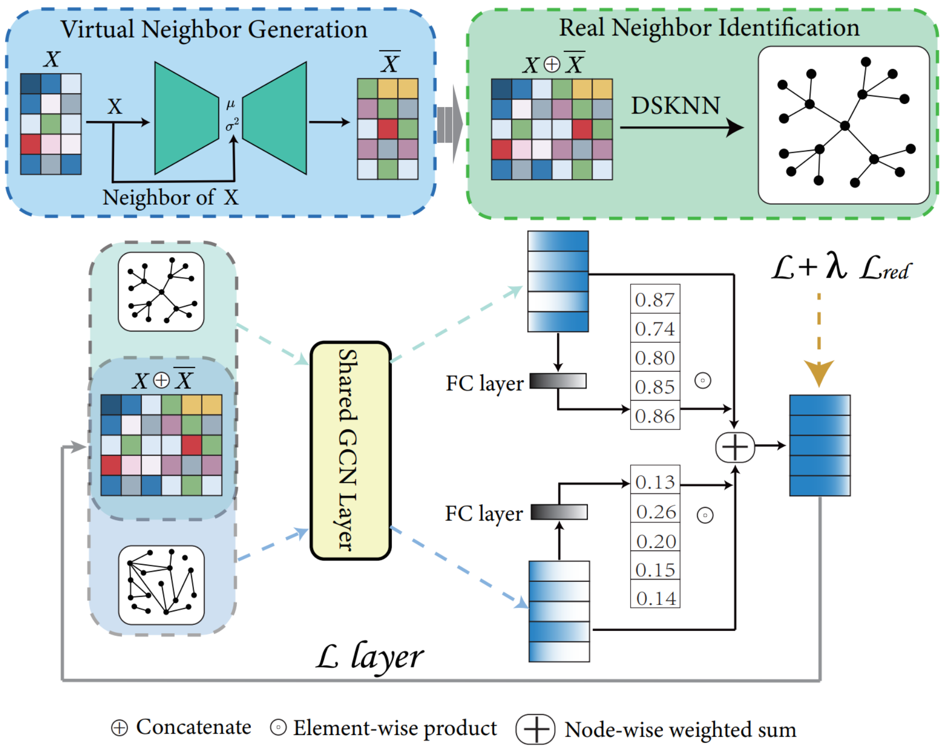
- DSKNN-graph
  - 1. Can reduce the probability of OOC nodes.
  - 2. Allowing flexible addition or removal of edges.

Number of UC nodes	Cora	Citeseer	Pubmed
Original Graph	633	485	708
DSKNN Graph	660	711	720
Combine Graph	833	840	879
Improve Ratio	31.6%	73.2%	24.2%

Solution: we just need to make sure that the number of edges in constructing the DSKNN graph is much larger than the average degree of the original graph.

University of Electronic Science and Technology of China

Overall Architecture



- Concatenate virtual neighbors'' feature as input feature:

$$\overline{\mathcal{X}} = \mathbf{X} \oplus \overline{\mathbf{X}}$$

- Propagating the features on the original graph and the DSKNN graph :

$$\mathbf{H}_{ori}^{(l)} = \hat{\mathbf{A}}\mathbf{H}^{(l-1)}\mathbf{W}^{(l-1)}, \mathbf{H}_{ds}^{(l)} = \hat{\mathbf{S}}\mathbf{H}^{(l-1)}\mathbf{W}^{(l-1)}$$

- Adaptive node-level assembling:

$$\mathbf{H}^{(l)} = \operatorname{diag}\left(\boldsymbol{\lambda}_0^{(l)}\right)\mathbf{H}_{ori}^{(l)} + \operatorname{diag}\left(\boldsymbol{\lambda}_1^{(l)}\right)\mathbf{H}_{ds}^{(l)}, \quad \boldsymbol{\lambda}_0^{(l)} + \boldsymbol{\lambda}_1^{(l)} = \mathbf{1}$$

$$\boldsymbol{\lambda}_0^{(l)} = \sigma\left(FC_0^{(l)}\left(\mathbf{H}_{ori}^{(l)}\right)\right), \boldsymbol{\lambda}_1^{(l)} = \sigma\left(FC_1^{(l)}\left(\mathbf{H}_{ds}^{(l)}\right)\right)$$

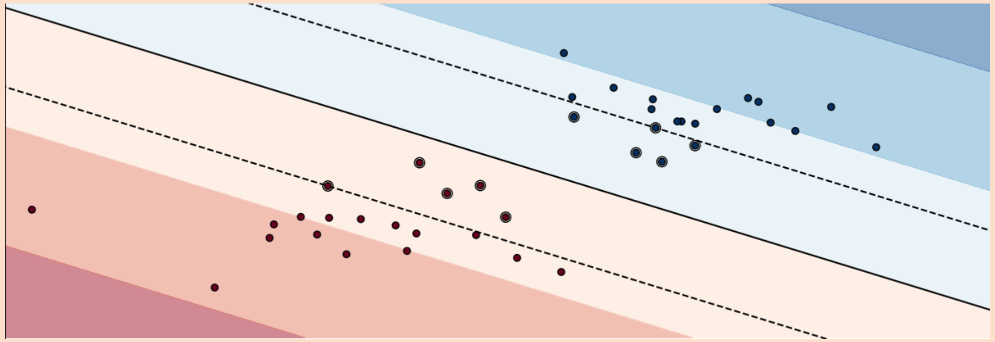
$$\left[\boldsymbol{\lambda}_0^{(l)}, \boldsymbol{\lambda}_1^{(l)}\right] = \frac{\left[\boldsymbol{\lambda}_0^{(l)}, \boldsymbol{\lambda}_1^{(l)}\right]}{\max\left(\left\|\left[\boldsymbol{\lambda}_0^{(l)}, \boldsymbol{\lambda}_1^{(l)}\right]\right\|_2, \epsilon\right)}$$

Method

Some Problem

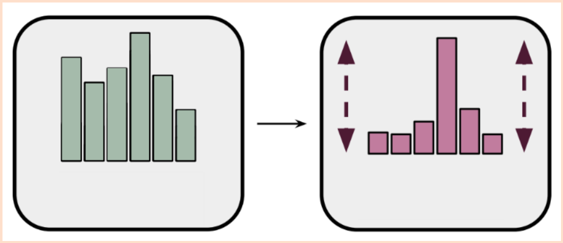
Fundamental Assumption in Semi-Supervised Learning

In semi-supervised learning, the classifier's decision boundary should avoid high-density regions of the data distribution.



An accomplish way

Ensuring the classifier outputs low-entropy predictions on unlabeled data.



In Adaptive node-level assembling — Violation

Lemma 1

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$$

- When the output layer assembles the logits, the entropy will increase beyond a linear combination of the two view.

Solution

Entropy Reduction Loss:

$$\mathcal{L}_{red} = \frac{1}{c}\sum_{i=1}^{c}\left(\mathbf{y}_i - \mathbf{y}_i^{\frac{1}{\tau}} \Big/ \sum_{j}^{c}\mathbf{y}_j^{\frac{1}{\tau}}\right)^2 + \mathbb{I}(||\mathbf{y}_{ori} - \mathbf{y}_{ds}||_2)$$

Experiments

University of Electronic Science and Technology of China

Main Results

Datasets	Cora	Citeseer	Pubmed	Computers	Photo	Physics	CS
GCN	81.5±0.82	70.9±0.71	79.0±0.52	82.6±2.43	91.2±1.21	92.8±1.00	91.1±0.52
GAT	83.0±0.41	71.1±0.51	79.1±0.44	78.0±19.0	85.7±20.3	92.5±0.94	90.5±0.61
APPNP	83.3±0.51	72.5±0.62	79.9±0.32	82.2±2.13	90.8±1.32	93.7±0.69	92.5±0.32
GCN-LPA	83.1±0.73	72.6±0.80	78.6±1.32	83.5±1.41	91.1±0.83	93.6±1.06	91.8±0.42
DAGNN	84.4±0.57	73.3±0.65	80.5±0.53	83.5±1.28	92.0±1.22	94.0±0.62	91.5±0.33
*w*GCN	83.1 ±0.31	73.9 ±0.46	80.8±0.25	83.6±0.86	92.4±0.18	92.8±0.23	89.3±0.14
AERO-GNN	83.9 ±0.51	73.2 ±0.68	80.6±0.55	83.3 ±0.72	91.1±0.83	93.3±0.65	92.0±0.71
Ours	**84.8**± **0.53**	**75.3** ± **0.41**	**81.7**±**0.88**	**84.0**±**1.25**	**92.9**±**0.56**	**94.3**±**0.25**	**93.4**±**0.18**

- In the node classification task, our proposed method outperformance the SOTA baseline.

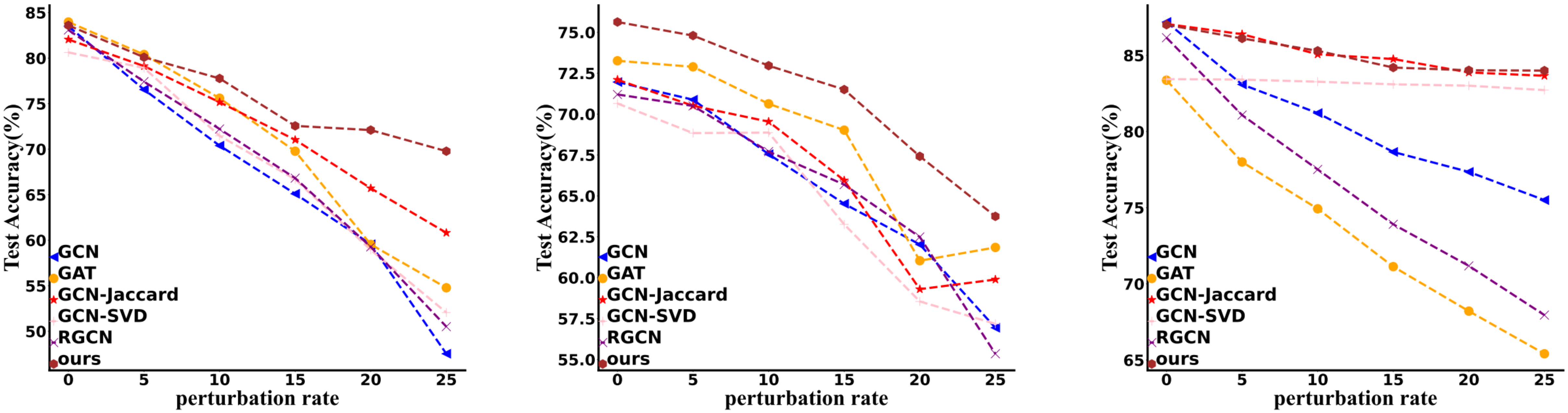
Datasets	Cora		Citeseer		Pubmed		Computers		Photo		Physics		CS	
	UC nodes	OOC nodes	UC nodes	OOC nodes	UC nodes	OOC nodes	UC nodes	OOC nodes	UC nodes	OOC nodes	UC nodes	OOC nodes	UC nodes	OOC nodes
GCN	87.01 ± 0.6	73.95 ± 1.1	77.75±0.5	62.66±0.8	83.66± 0.3	67.05±1.0	87.41±0.5	70.13±0.8	96.58±0.7	78.02±1.4	97.01±0.2	86.45±0.3	95.65±0.4	84.53± 0.6
APPNP	87.47± 0.5	76.21±1.3	78.21±0.6	67.59±0.9	84.36±0.5	67.96±1.1	87.23 ± 0.9	69.84± 2.6	95.98± 0.8	78.13±1.5	97.13± 0.5	89.25± 0.9	95.31± 0.2	87.01 ± 0.5
DAGNN	87.80 ±0.5	78.52±1.5	78.33±0.7	68.27±0.93	84.48±0.8	68.32±0.7	88.21±0.7	71.97± 1.5	95.36±0.8	80.64±1.2	97.16±0.5	89.98±0.7	94.75±0.2	87.53±0.7
AERO-GNN	87.74±0.3	77.38±0.8	78.14±0.8	68.78±1.0	85.38±0.3	69.79±1.1	88.56±0.8	71.72±1.3	96.34± 0.6	77.65± 1.0	97.03±0.4	88.65±0.9	95.89±0.6	86.01±1.1
Ours	87.70±0.5	**79.26**± **0.7**	78.39±0.5	**72.04**±**1.5**	85.54± 0.3	**73.16** ± **1.0**	88.12±0.8	**73.39**± **1.5**	95.57± 0.5	**82.75**±**0.8**	97.12±0.2	**91.15**±**0.3**	95.53±0.1	**89.51**±**0.3**

- Most methods (including ours) show similar effectiveness on UC nodes. The key factor differentiating their performance is their behavior on OOC nodes. Thus, research on GCNs should primarily focus on OOC nodes.

- Our proposed method significantly improves the performance of GCNs on OOC nodes.

Experiments

Adversarial Robustness-Metaattack

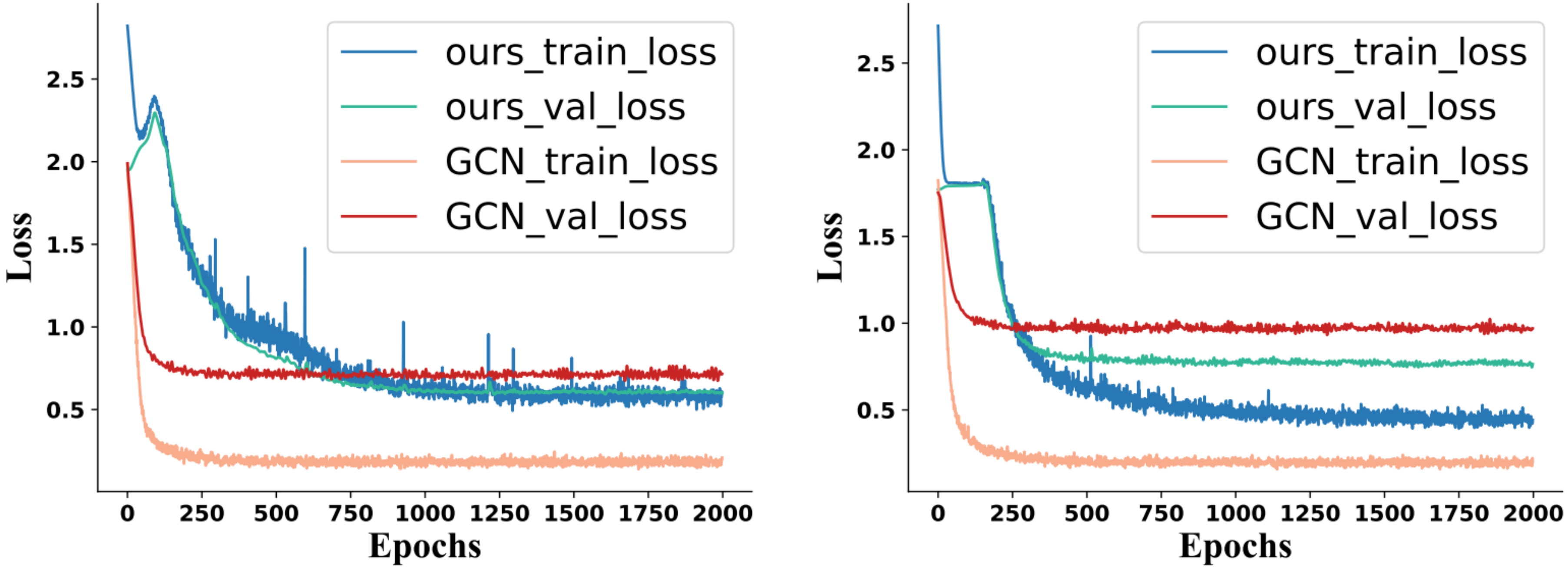


(a) Cora (b) Citeseer (c) Pubmed

- Our proposed method has strong adversarial robustness

Experiments

Analysis Generalization Ability

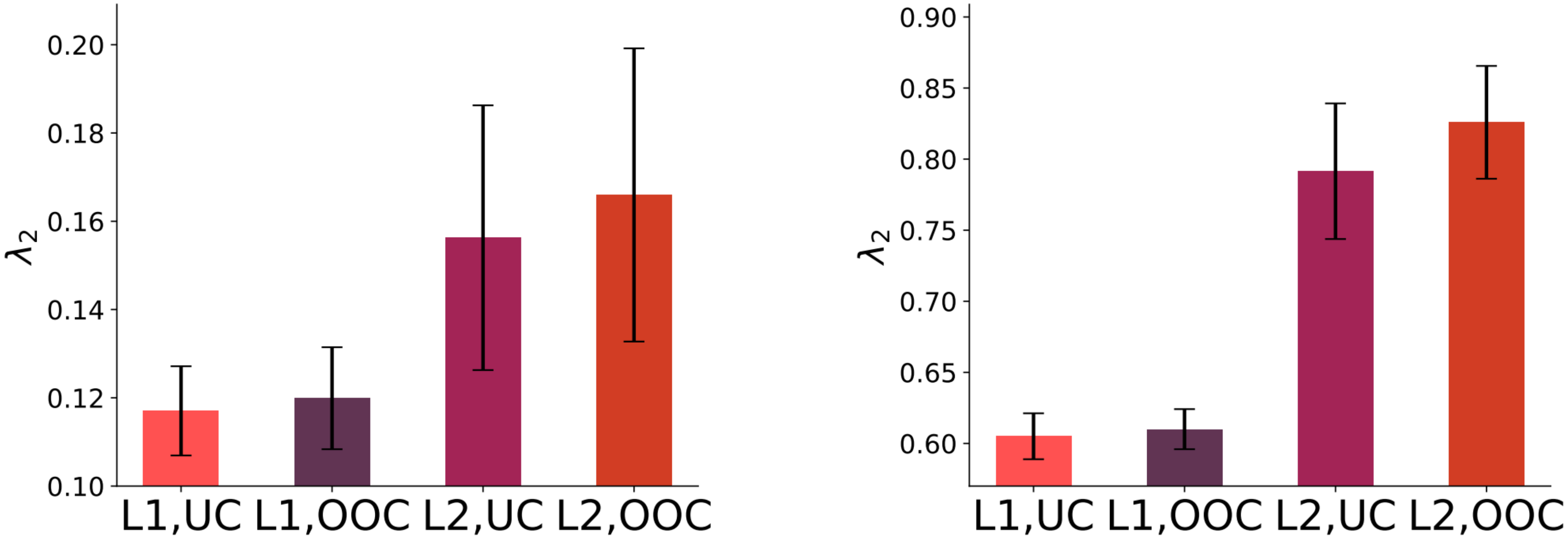


(a) Cora (b) Citeseer

- Our proposed improves the GCN's generalization ability.

Experiments

Analysis Adaptive Node-level Assembling



(a) Cora (b) Pubmed

- The OOC nodes have heavier average weights in the second layer of the DSKNN side compared to the UC nodes, suggesting greater benefit for OOC nodes from the DSKNN side.
- The weights learned by each layer are differentiated.

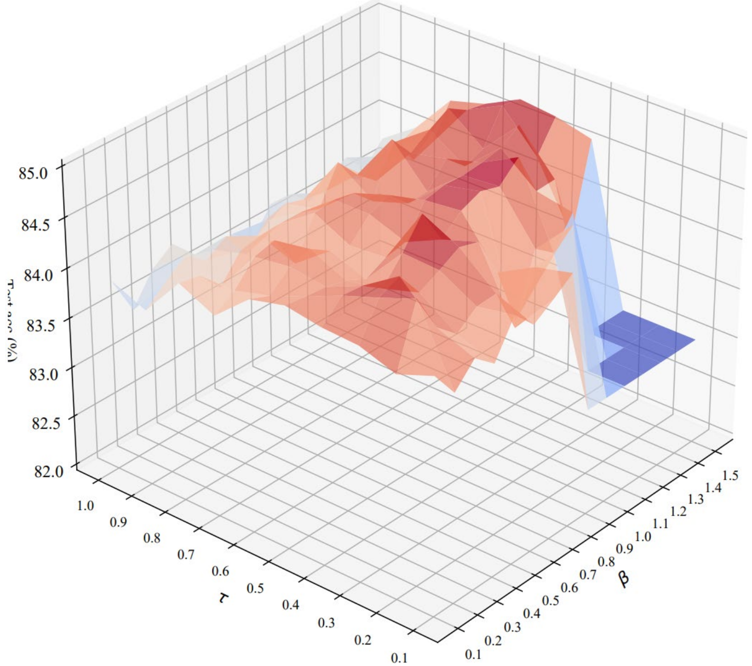
Experiments

Analysis Adaptive Node-level Assembling

Ablation	Cora	Citeseer	Pubmed
DaGCN	84.8±0.53	75.3±0.41	81.7±0.88
- w/o VNG	84.2±0.96	74.5±0.66	81.2±1.00
- w/o RNG	83.6±0.46	73.6±0.37	80.7±0.62
- w/o ERL	84.0±0.56	73.8±0.72	81.3±0.75
GCN	81.5±0.82	70.9±0.71	79.0±0.52

- All components are valid.
- The DSKNN-graph part played the biggest effect.

Analysis Parameter Sensitivity



- the temperature parameter τ is significantly important, since when τ is in the interval [0.4, 0.8], the model performance maintains an excellent level. The DSKNN-graph part played the biggest effect.
- if we ensure that τ is in a suitable range, the selection of β is not sensitive.

Summary

Conclusion

- vanilla GCN has been able to achieve high-quality representation learning on UC nodes. The advanced model should focus on improving OOC nodes to promote GCN.
- We provide algorithms for locating OOC nodes and provide directions and models to promote OOC nodes.

Future Work

- Optimize graph structure from the perspective of reducing OOC nodes and Generalization.

电子科技大学
University of Electronic Science and Technology of China

Thanks!