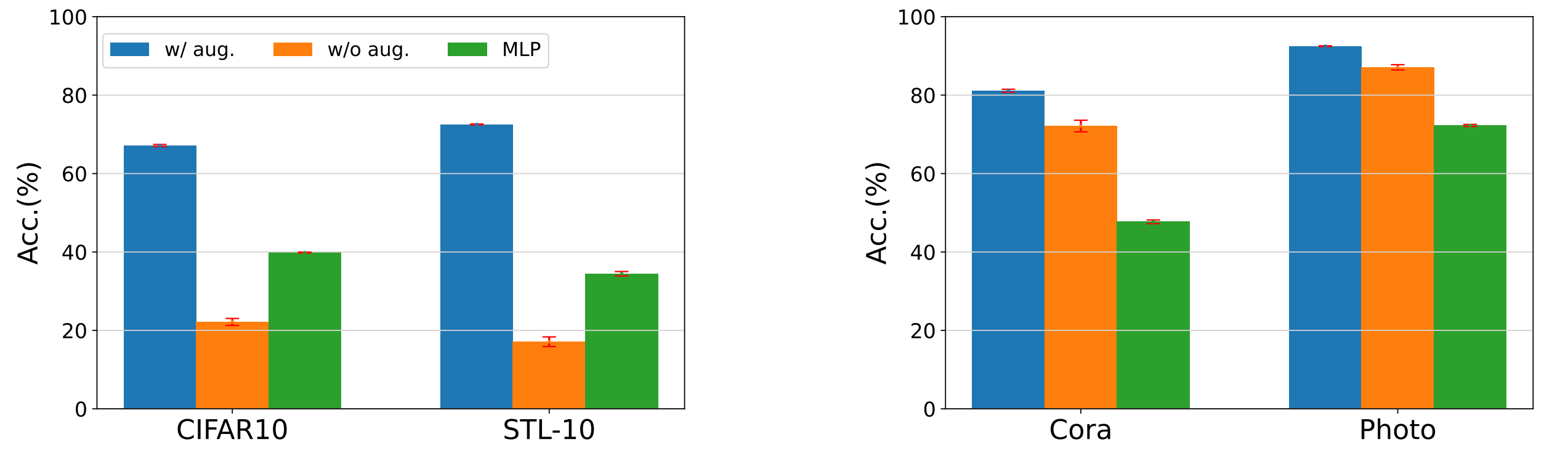
HomoGCL: Rethinking Homophily in Graph Contrastive Learning

Wen-Zhi Li [1,2] Chang-Dong Wang [1] Hui Xiong [2,3] Jian-Huang Lai [1]

[1]CSE, Sun Yat-sen University [2]AI Thrust, HKUST (GZ) [3]CSE, HKUST

Background & Motivation



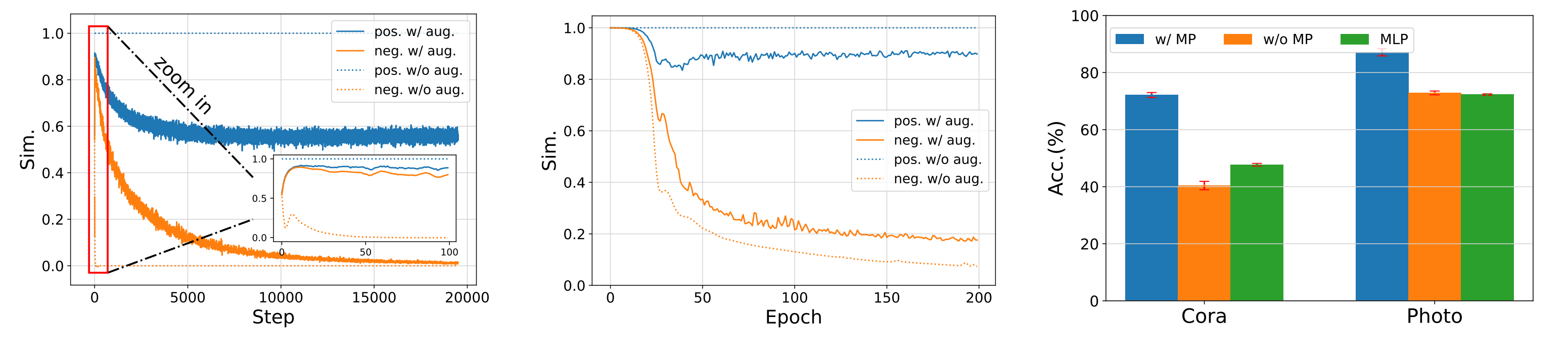
(a) Vision datasets with SimCLR (b) Graph datasets with GRACE

Figure 1. Performance of CL in vision and graph domains with/without augmentation.

- CL: the "augmenting-contrasting" paradigm, where the similarity between two augmentations of a sample (positive pair) is maximized while the similarities between other samples (negative pairs) are minimized.
- **Empirical observation**: GCL without augmentation can also achieve decent performance, which is quite different from VCL.

What causes the huge gap between the performance declines of GCL and VCL when data augmentation is not leveraged? 😲

Empirical Study



(a) Similarity histogram on CIFAR10 (b) Similarity histogram on Cora (c) Ablation study on Cora and Photo

Figure 2. Empirical study on graph homophily.

# For (a), (b)

- **Obs #1**: The similarity between negative pairs drops to 0 swiftly on CIFAR10 w/o augmentation.
- **Obs #2**: The similarity between negative pairs drops gradually on Cora w/o augmentation, which is analogous to its counterpart with augmentation.

Hypothesis: Message passing in GNN enables information aggregation from neighbors, which leverages homophily implicitly to avoid trivial discrimination.

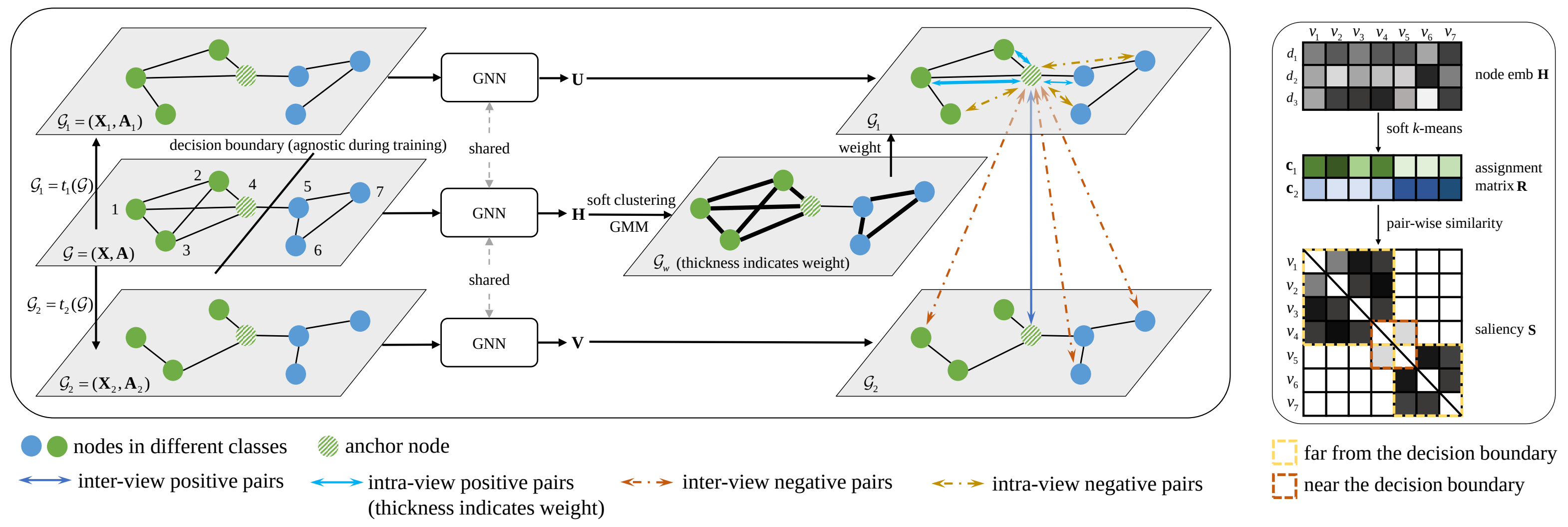
# For (c)

- **Obs #1**: GRACE (w/o MP) is only on par with or even worse than MLP.
- **Obs #2**: GRACE (w/ MP) outperforms w/o MP and MLP by a large margin.

Analysis: Nodes in GRACE (w/o MP) cannot propagate features to their neighbors, which degenerates them to a similar situation of VCL w/o augmentation. GRACE (w/ MP) can still maintain the performance even without raw features.

Conclusion: Message passing which relies on the homophily assumption is the key factor of GCL.

HomoGCL: Leveraging Graph Homophily Explicitly


$\mathcal{G}_1=(\mathbf{X}_1,\mathbf{A}_1)$
decision boundary (agnostic during training)
$\mathcal{G}_1=t_1(\mathcal{G})$
$\mathcal{G}=(\mathbf{X},\mathbf{A})$
$\mathcal{G}_2=t_2(\mathcal{G})$
$\mathcal{G}_2=(\mathbf{X}_2,\mathbf{A}_2)$
1
2
3
4
5
6
7
GNN
shared
GNN
shared
GNN
U
H
V
soft clustering
GMM
$\mathcal{G}_w$ (thickness indicates weight)
weight
$\mathcal{G}_1$
$\mathcal{G}_2$
nodes in different classes
anchor node
inter-view positive pairs
intra-view positive pairs (thickness indicates weight)
inter-view negative pairs
intra-view negative pairs
v_1 v_2 v_3 v_4 v_5 v_6 v_7
d_1 d_2 d_3
node emb H
soft k-means
c_1 c_2
assignment matrix R
pair-wise similarity
saliency S
far from the decision boundary
near the decision boundary

(a) Workflow of HomoGCL (b) Soft clustering

Figure 3. The pipeline of HomoGCL.

# Challenges

- As inter-class edges exist near the decision boundary between two classes, simply assigning neighbor nodes as positive is non-ideal (i.e., *false positive*).
- **Estimating the probability** of neighbor nodes being positive **in an unsupervised manner**.

# Soft clustering for pair-wise node similarity

- Hard k-means → Soft k-means: treating k-means as a special case of GMM.

$$p\left(\boldsymbol{c}_j \mid \boldsymbol{h}_i\right)=\frac{p\left(\boldsymbol{c}_j\right) p\left(\boldsymbol{h}_i \mid \boldsymbol{c}_j\right)}{\sum_{r=1}^{k} p\left(\boldsymbol{c}_r\right) p\left(\boldsymbol{h}_i \mid \boldsymbol{c}_r\right)},$$

with $p\left(\boldsymbol{h}_i \mid \boldsymbol{c}_j\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\|\boldsymbol{h}_i-\boldsymbol{c}_j\|_2}{2\sigma^2}\right)$ and $p(\boldsymbol{c}_1)=p(\boldsymbol{c}_2)=\cdots=p(\boldsymbol{c}_k)$.

- Cluster assignment matrix $\mathbf{R}_{ij}=p(\boldsymbol{c}_j \mid \boldsymbol{h}_i)$ indicating the soft clustering value between node v_i and cluster $\boldsymbol{c}_j$.
- Saliency $\mathbf{S}_{ij}=\operatorname{norm}(\mathbf{R}_i)\cdot\operatorname{norm}(\mathbf{R}_j^\top)$ indicating the connection intensity between v_i and v_j, which is an estimated probability being true positive.

# Positive set

$$\text{pos}=\underbrace{e^{\theta(\boldsymbol{u}_i,\boldsymbol{v}_i)/\tau}}_{\text{inter-view positive pair}}+\underbrace{\sum_{j\in\mathcal{N}_{\boldsymbol{u}}(i)} e^{\theta(\boldsymbol{u}_i,\boldsymbol{u}_j)/\tau}\cdot\mathbf{S}_{ij}}_{\text{intra-view positive pairs}}.$$

# Homophily loss

$$\mathcal{L}_{homo}=\frac{1}{k|\mathcal{E}|}\sum_{r=1}^{k}\sum_{(v_i,v_j)\in\mathcal{E}}\operatorname{MSE}\left(p\left(\boldsymbol{c}_r \mid \boldsymbol{h}_i\right), p\left(\boldsymbol{c}_r \mid \boldsymbol{h}_j\right)\right).$$

# Theoretical insight

The newly proposed positive set in the contrastive loss of HomoGCL introduces a stricter lower bound of MI between raw node features $\mathbf{X}$ *and node embeddings* $\mathbf{U}$ *and* $\mathbf{V}$ *in two augmented views, comparing with the raw contrastive loss proposed by GRACE. Formally,*

$$\mathcal{L}_{HomoGCL}\leq\mathcal{L}_{GRACE}\leq I(\mathbf{X};\mathbf{U},\mathbf{V}).$$

Experiments

Table 1. Node classification (accuracy(%) ±std). $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{Y}$ correspond to node features, graph adjacency matrix, and node labels respectively. "↑" and "↓" refer to performance improvement and drop compared with the same GRACE base model.

Model	Training Data	Cora	CiteSeer	PubMed	Photo	Computer
Raw features	**X**, **Y**	47.7±0.4	46.5±0.4	71.4±0.2	72.27±0.00	73.81±0.00
DeepWalk	**A**	70.7±0.6	51.4±0.5	74.3±0.9	89.44±0.11	85.68±0.06
Node2Vec	**A**	70.1±0.4	49.8±0.3	69.8±0.7	87.76±0.10	84.39±0.08
GCN	**X**, **A**, **Y**	81.5±0.4	70.2±0.4	79.0±0.2	92.42±0.22	86.51±0.54
GAT	**X**, **A**, **Y**	83.0±0.7	72.5±0.7	79.0±0.3	92.56±0.35	86.93±0.29
GAE	**X**, **A**	71.5±0.4	65.8±0.4	72.1±0.5	91.62±0.13	85.27±0.19
VGAE	**X**, **A**	73.0±0.3	68.3±0.4	75.8±0.2	92.20±0.11	86.37±0.21
DGI	**X**, **A**	82.3±0.6	71.8±0.7	76.8±0.6	91.61±0.22	83.95±0.47
GMI	**X**, **A**	83.0±0.3	72.4±0.1	79.9±0.2	90.68±0.17	82.21±0.31
InfoGCL	**X**, **A**	83.5±0.3	**73.5**±0.4	79.1±0.2	-	-
MVGRL	**X**, **A**	83.5±0.4	73.3±0.5	80.1±0.7	91.74±0.07	87.52±0.11
BGRL	**X**, **A**	82.7±0.6	71.1±0.8	79.6±0.5	92.80±0.08	88.23±0.11
AFGRL	**X**, **A**	79.8±0.2	69.4±0.2	80.0±0.1	92.71±0.23	88.12±0.27
COSTA	**X**, **A**	82.2±0.2	70.7±0.5	80.4±0.3	92.43±0.38	88.37±0.22
CCA-SSG	**X**, **A**	84.0±0.4	73.1±0.3	81.0±0.4	92.84±0.18	88.27±0.32
GRACE	**X**, **A**	81.5±0.3	70.6±0.5	80.2±0.3	92.15±0.24	86.25±0.25
GCA	**X**, **A**	81.4±0.3(↓0.1)	70.4±0.4(↓0.2)	80.7±0.5(↑0.5)	92.53±0.16(↑0.38)	87.80±0.23(↑1.55)
ProGCL	**X**, **A**	81.2±0.4(↓0.3)	69.8±0.5(↓0.8)	79.2±0.2(↓1.0)	92.39±0.11(↑0.24)	87.43±0.21(↑1.18)
ARIEL	**X**, **A**	83.0±1.3(↑1.5)	71.1±0.9(↑0.5)	74.2±0.8(↓6.0)	91.80±0.24(↓0.35)	87.07±0.33(↑0.82)
HomoGCL	**X**, **A**	**84.5**±0.5(↑3.0)	72.3±0.7(↑1.7)	**81.1**±0.3(↑0.9)	**92.92**±0.18(↑0.77)	**88.46**±0.20(↑2.21)

[1] The results not reported are due to unavailable code.

Table 2. Node clustering.

Dataset	Photo		Computer	
Metric	NMI	ARI	NMI	ARI
GAE	0.616±0.01	0.494±0.01	0.441±0.00	0.258±0.00
VGAE	0.530±0.04	0.373±0.04	0.423±0.00	0.238±0.00
DGI	0.376±0.03	0.264±0.03	0.318±0.02	0.165±0.02
HDI	0.429±0.01	0.307±0.01	0.347±0.01	0.216±0.06
MVGRL	0.344±0.04	0.239±0.04	0.244±0.00	0.141±0.00
BGRL	0.668±0.03	0.547±0.04	0.484±0.00	0.295±0.00
AFGRL	0.618±0.01	0.497±0.03	0.478±0.03	0.334±0.04
GCA	0.614±0.00	0.494±0.00	0.426±0.00	0.246±0.00
gCooL	0.632±0.00	0.524±0.00	0.474±0.02	0.277±0.02
HomoGCL	**0.671**±0.02	**0.587**±0.02	**0.534**±0.00	**0.396**±0.00

Table 3. HomoGCL + BGRL.

Model	PubMed	Photo	Computer
BGRL	79.6	92.80	88.23
+**HomoGCL**	80.8(↑1.2)	93.53(↑0.73)	90.01(↑1.79)

Table 4. Node classification ogbn-arXiv.

Model	Validation	Test
MLP	57.65±0.12	55.50±0.23
node2vec	71.29±0.13	70.07±0.13
GCN	73.00±0.17	71.74±0.29
GraphSAGE	72.77±0.16	71.49±0.27
Random-Init	69.90±0.11	68.94±0.15
DGI	71.26±0.11	70.34±0.16
G-BT	71.16±0.14	70.12±0.18
GRACE full-graph	OOM	OOM
GRACE-Subsampling (k=2)	60.49±3.72	60.24±4.06
GRACE-Subsampling (k=8)	71.30±0.17	70.33±0.18
GRACE-Subsampling (k=2048)	72.61±0.15	71.51±0.11
ProGCL	72.45±0.21	72.18±0.09
BGRL	72.53±0.09	71.64±0.12
HomoGCL	**72.85**±0.10	**72.22**±0.15

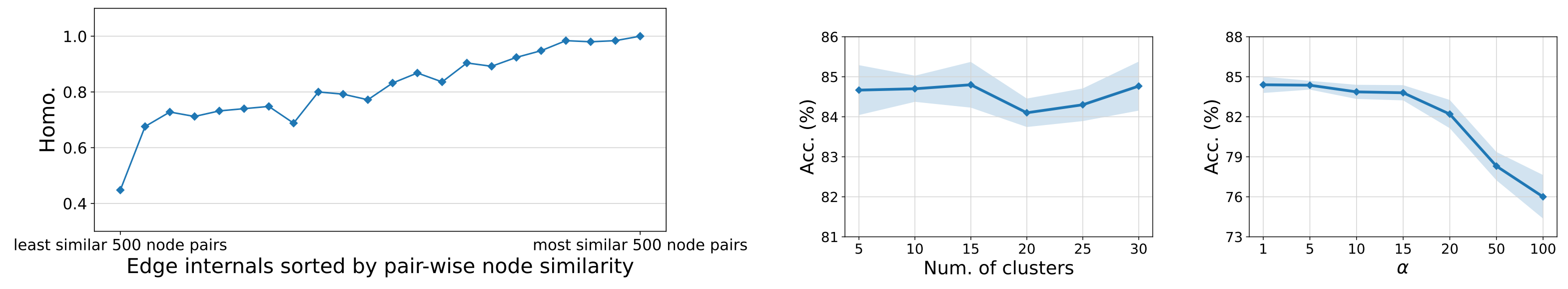
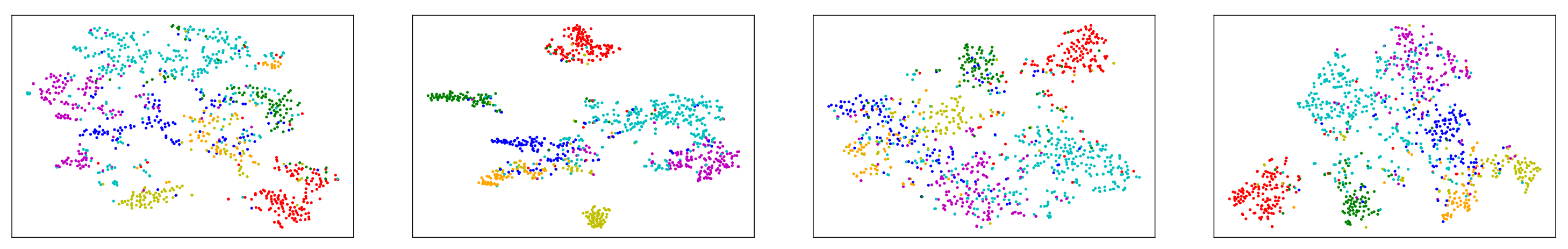


Figure 4. **(Left)**: The saliency $\mathbf{S}$ can effectively estimate the probability of neighbor nodes being positive as more salient edges (more similar node pairs) tend to have larger homophily. **(Middle & Right)**: Hyperparameter analysis on the number of clusters and weight coefficient α. All on Cora dataset.



(a) GRACE (b) GRACE + HomoGCL (c) BGRL (d) BGRL + HomoGCL

Figure 5. Visualization of node embeddings on Cora via t-SNE. Each node is colored by its label.

Code: https://github.com/wenzhilics/HomoGCL Paper: https://arxiv.org/abs/2306.09614 Contact: liwzh63@mail2.sysu.edu.cn