NEURAL INFORMATION PROCESSING SYSTEMS

Uncovering the Structural Fairness in Graph Contrastive Learning

Ruijia Wang[1], Xiao Wang[1], Chuan Shi[1], Le Song[2]

[1]Beijing University of Posts and Telecommunications

[2]BioMap and MBZUAI

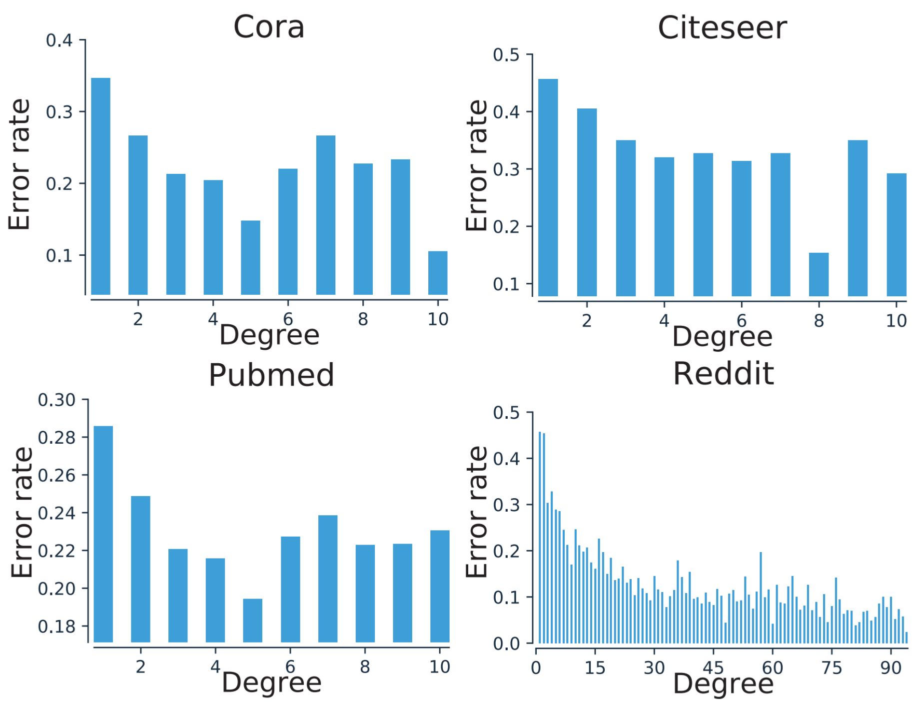
NEURAL INFORMATION PROCESSING SYSTEMS

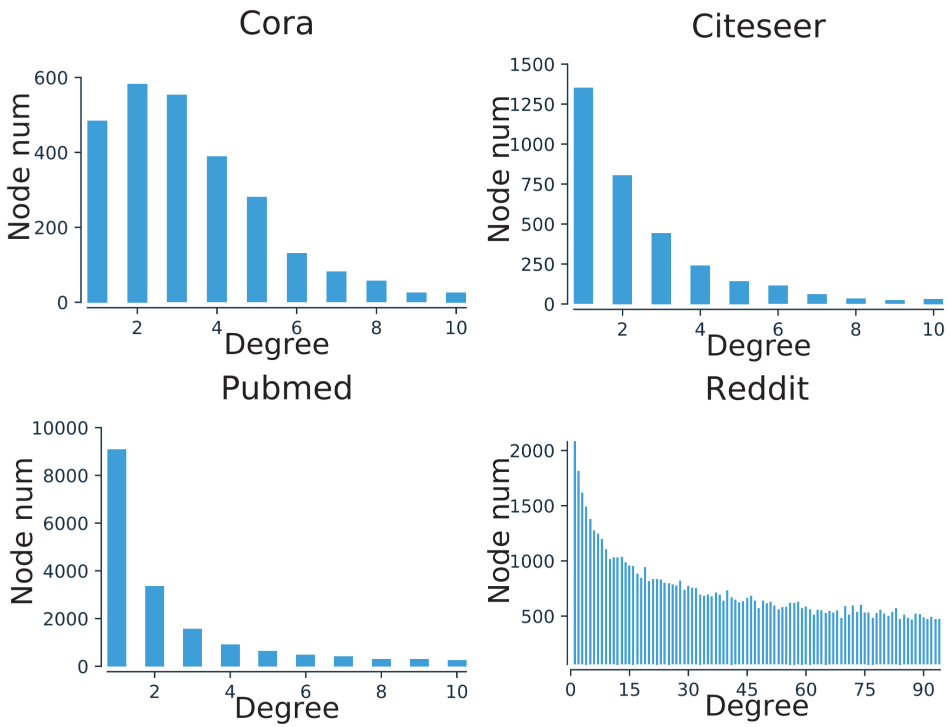
CONTENTS

Background

Graph Convolutional Network (GCN[1])

- Node degrees of real-world graphs often follow a **long-tailed distribution**.

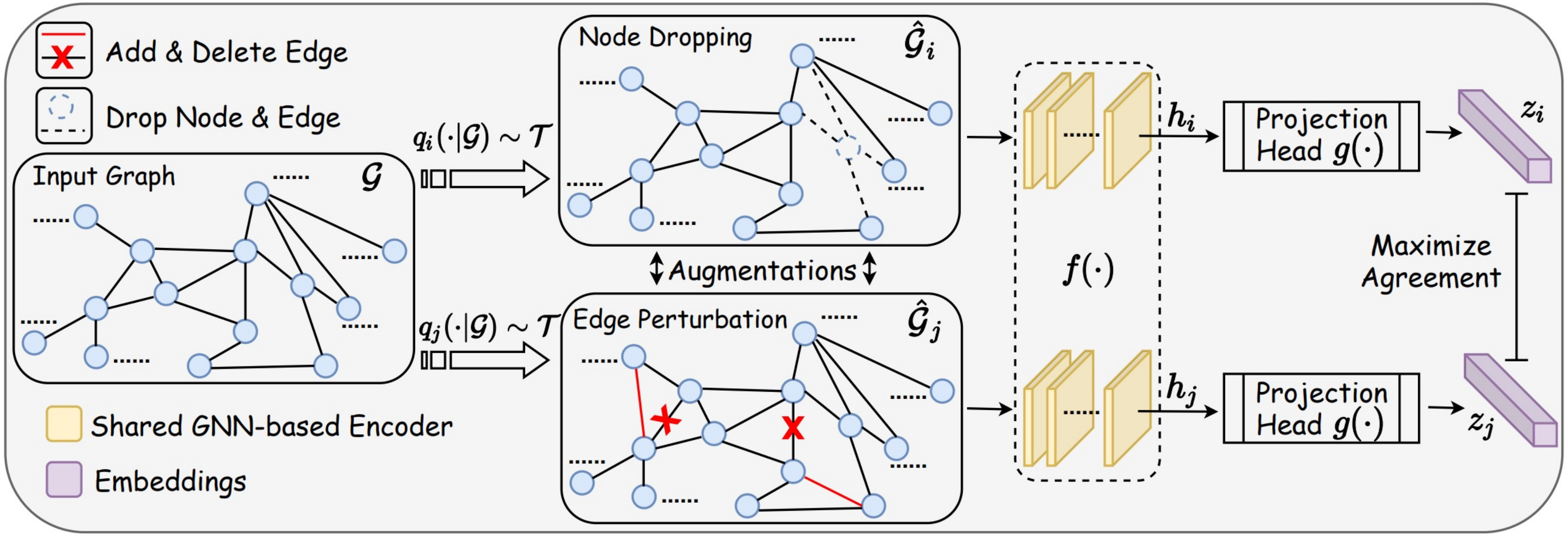




GCN exhibits a structural unfairness.

[1] T. N. Kipf, and M. Welling. Semi-supervised Classification with Graph Convolutional Networks. ICLR 2017.

Graph Contrastive Learning (GCL)

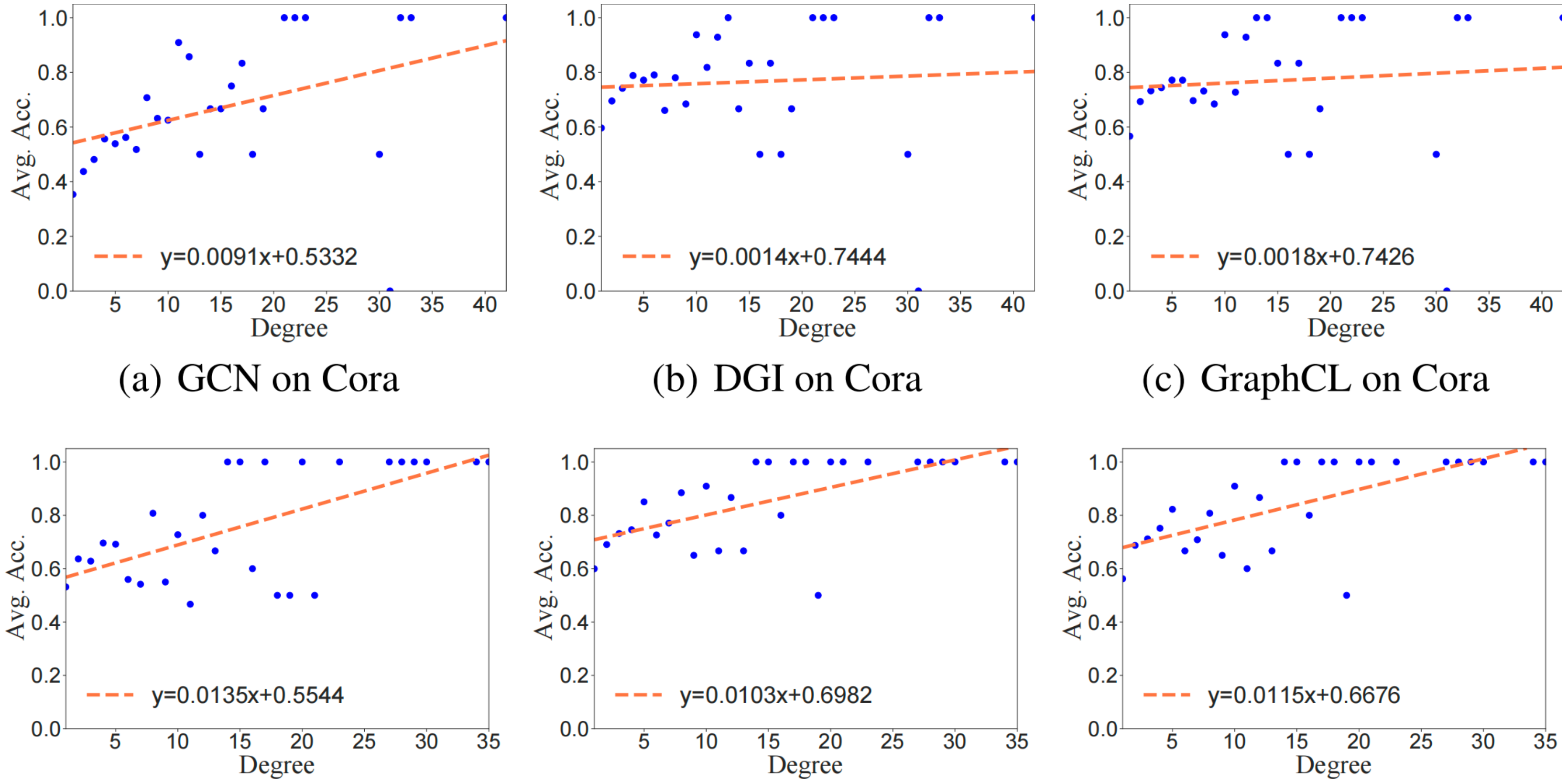


- GCL integrates the power of GCN and contrastive learning.
- GCL relieves GCN from annotations, and displays SOTA performance in many tasks.

Will GCL present the same structure unfairness as GCN?

Investigation and Analysis

Exploring the Behavior of Graph Contrastive Learning on Degree Bias



(a) GCN on Cora (b) DGI on Cora (c) GraphCL on Cora

(d) GCN on Citeseer (e) DGI on Citeseer (f) GraphCL on Citeseer

- A **smaller performance gap** exists in GCL methods than that of GCN.

Why is graph contrastive learning fairer to degree bias?

Analysis on the Structural Fairness of Graph Contrastive Learning

Preliminary Notations

- Let $G=(\mathcal{V},\mathcal{E},X)$ be a graph, $X=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]\in\mathbb{R}^{N\times B}$ is node feature matrix.
- The edges can be represented by an adjacency matrix $A\in\{0,1\}^{N\times N}$.
- Assume **the augmentation set** $\mathcal{T}$ consisting of all transformations on topology.
- **Positive samples** generated from ego network $\mathcal{G}_i$ of node v_i denoted as $\mathcal{T}(\mathcal{G}_i)$.
- Here we focus on **topological augmentation and single-layer GCN**

$$f(\mathcal{G}_i)=ReLU(\tilde{L}_i XW) \qquad \tilde{L}=\tilde{D}^{-1}\tilde{A},\ \tilde{A}=A+I,\ \tilde{D}_{ii}=\sum_j \tilde{A}_{ij}$$

- We consider a **community indicator** F_f

$$F_f(\mathcal{G}_i)=\underset{k\in[K]}{\arg\min}\|f(\mathcal{G}_i)-\mu_k\| \qquad \mu_k=\mathbb{E}_{v_i\in C_k}\mathbb{E}_{\hat{\mathcal{G}}_i\in\mathcal{T}(\mathcal{G}_i)}[f(\hat{\mathcal{G}}_i)]$$

- The **error** of community indicator can be formulated as

$$\mathrm{Err}(F_f)=\sum_{k=1}^{K}\mathbb{P}[F_f(\mathcal{G}_i)\neq k,\forall v_i\in C_k]$$

- we denote $S_\varepsilon=\{v_i\in\cup_{k=1}^{K}C_k:\forall\hat{\mathcal{G}}_i^1,\hat{\mathcal{G}}_i^2\in\mathcal{T}(\mathcal{G}_i),\|f(\hat{\mathcal{G}}_i^1)-f(\hat{\mathcal{G}}_i^2)\|\leq\varepsilon\}$ as **nodes with ε-close representations among graph augmentations**.

Analysis on the Structural Fairness of Graph Contrastive Learning

Theoretical Analysis

- Assume the **nonlinear transformation has M-Lipschitz continuity**

$$\|f(\mathcal{G}_i) - f(\mathcal{G}_j)\| = \|ReLU(\tilde{L}_i XW) - ReLU(\tilde{L}_j XW)\| \leq M\|\tilde{L}_i X - \tilde{L}_j X\|$$

- Graph augmentations are **uniformly sampled** with m augmented edges

$$\mathbb{P}[\hat{\mathcal{G}}_i = \mathcal{T}(\mathcal{G}_i)] = 1/C(N-1, m)$$

- Let there be a ball of **radius** βm such that for any augmentation

$$\|\tilde{L}_i X - \hat{L}_i X\|^2 \leq \beta m$$

Theorem 1 ***Intra-community Concentration.*** *Let pre-transformation representations $\tilde{L}X$ be sub-Gaussian random variable with variance σ^2. For all nodes $v_i \in S_\varepsilon$, if $\varepsilon^2 \leq \frac{\beta m}{6M^2\kappa}$, their representations $f(\mathcal{G}_i)$ fit sub-Gaussian distribution with variance $\sigma^2_{f,\varepsilon} \leq \frac{1}{\kappa}\sigma^2$ with $\kappa \geq 1$ where κ is a coefficient that reflects the degree of concentration.*

Relation between the intra-community concentration and the alignment of positive pairs

Analysis on the Structural Fairness of Graph Contrastive Learning

Theoretical Analysis

- Define the **augmentation distance** between nodes as the minimum distance between their pre-transformation representations

$$d_{\mathcal{T}}(v_i, v_j) = \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \|\hat{L}_i X - \hat{L}_j X\| = \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \|(\frac{\hat{A}_i}{\hat{d}_i} - \frac{\hat{A}_j}{\hat{d}_j})X\|$$

- Introduce **the definition of $(\alpha, \gamma, \hat{d})$ -augmentation** to measure the concentration of pre-transformation representations

Definition 1 $(\alpha, \gamma, \hat{d})$***-Augmentation.*** *The augmentation set $\mathcal{T}$ is a $(\alpha, \gamma, \hat{d})$-augmentation, if for each community C_k, there exists a subset $C_k^0 \subset C_k$ such that the following two conditions hold*

1. $\mathbb{P}[v_i \in C_k^0] \geq \alpha \mathbb{P}[v_i \in C_k]$ *where* $\alpha \in (0, 1]$,
2. $\sup_{v_i, v_j \in C_k^0} d_{\mathcal{T}}(v_i, v_j) \leq \gamma (\frac{B}{\hat{d}_{\min}^k})^{\frac{1}{2}}$ *where* $\gamma \in (0, 1]$,

where $\hat{d}_{\min}^k = \min_{v_i \in C_k^0, \hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i)} \hat{d}_i$, *and B is the feature dimension.*

Analysis on the Structural Fairness of Graph Contrastive Learning

Theoretical Analysis

- Assume that the representation is **normalized** by $\|f(\mathcal{G}_i)\| = r$ and let $p_k = \mathbb{P}[v_i \in C_k]$
- **Bound the inter-community distance and the error of the community indicator**

Theorem 2 ***Inter-community Scatter.*** *For a $(\alpha, \gamma, \hat{d})$-augmentation, if*

$$\mu_\ell^\top \mu_k < r^2\Big(1 - \rho_{\max}(\alpha,\gamma,\hat{d},\varepsilon) - \sqrt{2\rho_{\max}(\alpha,\gamma,\hat{d},\varepsilon)} - \frac{\Delta_\mu}{2}\Big) \tag{5}$$

holds for any pair of (ℓ, k) with $\ell \neq k$, then the error of the community indicator F_f can be bounded by $(1-\alpha) + R_\varepsilon$, where $\rho_{\max}(\alpha,\gamma,\hat{d},\varepsilon) = 2(1-\alpha) + \max_\ell\left(\frac{2R_\varepsilon}{p_\ell} + \frac{M\alpha\gamma\sqrt{B}}{r\sqrt{\hat{d}^\ell_{\min}}}\right) + \frac{2\alpha\varepsilon}{r})$ and $\Delta_\mu = 1 - \min_{k\in[K]}\|\mu_k\|^2/r^2$.

2 concentration of representations

Theorem 3 *The term R_ε is upper bounded by*

1 alignment of positive pairs

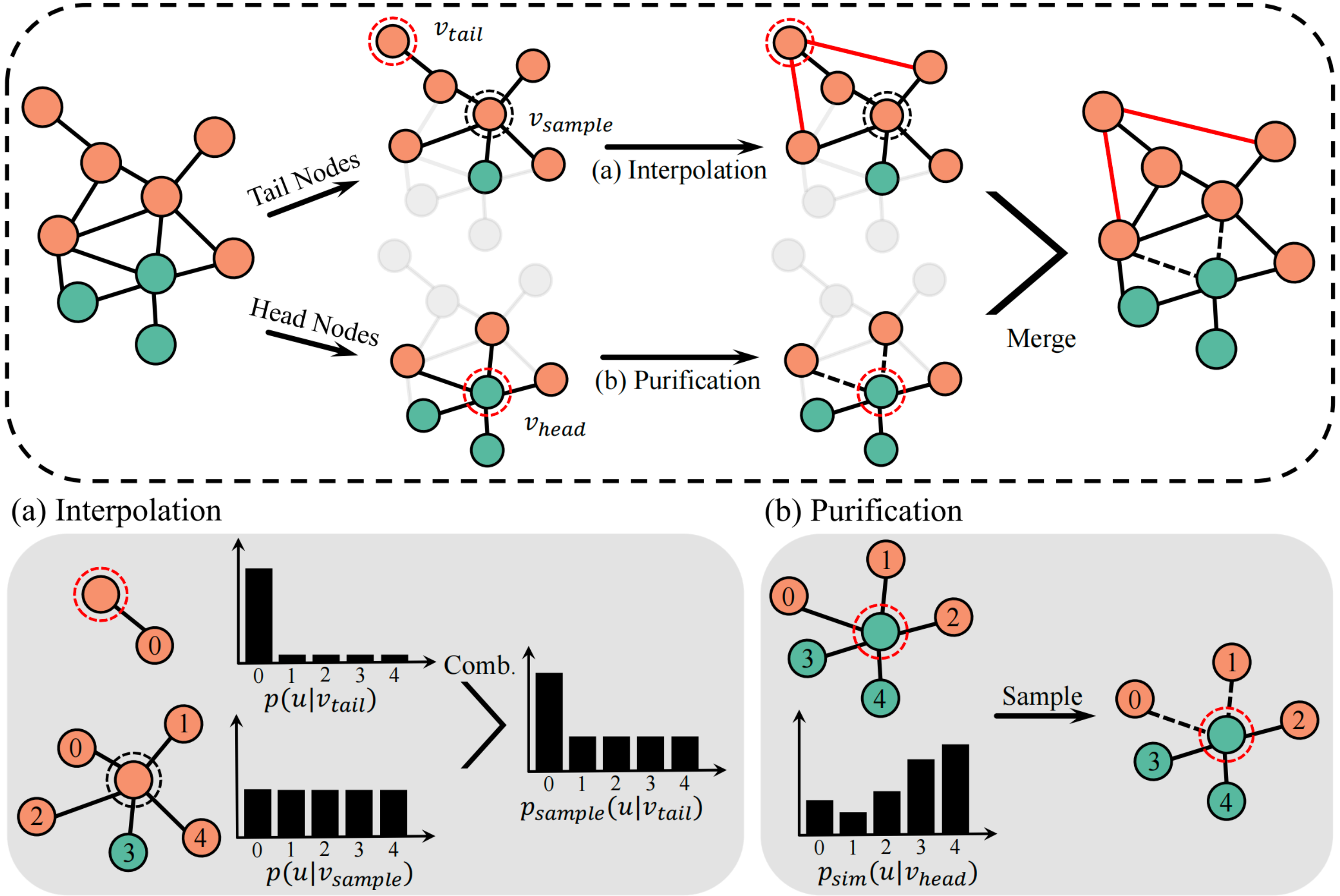
$$R_\varepsilon \leq \frac{[C(N-1,m)]^2}{\varepsilon}\mathbb{E}_{v_i}\mathbb{E}_{\hat{\mathcal{G}}_i^1,\hat{\mathcal{G}}_i^2\in\mathcal{T}(\mathcal{G}_i)}\|f(\hat{\mathcal{G}}_i^1) - f(\hat{\mathcal{G}}_i^2)\|. \tag{6}$$

GCL conform to a clearer community structure

The Proposed Model

Overview

Aim to increase intra-community edges while decreasing inter-community edges



Tail nodes

- **Interpolate the ego network** of the anchor tail node with that of a similar node.

Head nodes

- **Purify the neighborhood** by similarity-based sampling.

Graph Augmentation

Topology Augmentation

- We build the similarity matrix S based on **cosine similarity** of representations

$$S_{ij} = \mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j) \text{ for } i \neq j \text{ and } S_{ii} = 0 \text{ otherwise}$$

- For any tail node v_{tail} , we sample a node v_{sample} from the distribution $\mathrm{Multi}(\boldsymbol{s}_{tail})$.
- The similarity $\mathrm{sim}(\boldsymbol{h}_{tail}, \boldsymbol{h}_{sample})$ is used as the **interpolation** ratio ϕ

$$p_{sample}(u|v_{tail}) = \phi p(u|v_{tail}) + (1-\phi)p(u|v_{sample})$$

- For each head node v_{head} , we define the **similarity distribution for purification**

$$p_{sim}(u|v) = \mathrm{sim}(\boldsymbol{h}_u, \boldsymbol{h}_v) \text{ if } u \in \mathcal{N}(v) \text{ and } p(u|v) = 0$$

- We sample $d_{head}(1-p_{edr})$ neighbors without replacement.

Feature Augmentation

- We randomly sample a **mask** $\boldsymbol{m} \in \{0,1\}^B$ from a Bernoulli distribution $\mathrm{Ber}(1-p_{fdr})$

$$\hat{X} = [\boldsymbol{x}_1 \circ \boldsymbol{m}, \boldsymbol{x}_2 \circ \boldsymbol{m}, \cdots, \boldsymbol{x}_N \circ \boldsymbol{m}]$$

Optimization Objective

- Node representations $\boldsymbol{h}_i$ and $\boldsymbol{o}_i$ from different graph augmentations form the positive pair.
- Node representations of other nodes in graph augmentations are regarded as negative pairs.
- We define the **pairwise objective for each positive pair**[1] $(\boldsymbol{h}_i, \boldsymbol{o}_i)$ as

$$\ell(\boldsymbol{h}_i, \boldsymbol{o}_i) = \log \frac{e^{\theta(\boldsymbol{h}_i, \boldsymbol{o}_i)/\tau}}{e^{\theta(\boldsymbol{h}_i, \boldsymbol{o}_i)/\tau} + \sum_{k \neq i} e^{\theta(\boldsymbol{h}_i, \boldsymbol{o}_k)/\tau} + \sum_{k \neq i} e^{\theta(\boldsymbol{h}_i, \boldsymbol{h}_k)/\tau}}$$

where τ is a temperature parameter, and the critic $\theta(\boldsymbol{h}, \boldsymbol{o})$ is defined a $\mathrm{sim}(g(\boldsymbol{h}), g(\boldsymbol{o}))$.

- The overall objective to be maximized is **the average of all positive pairs**

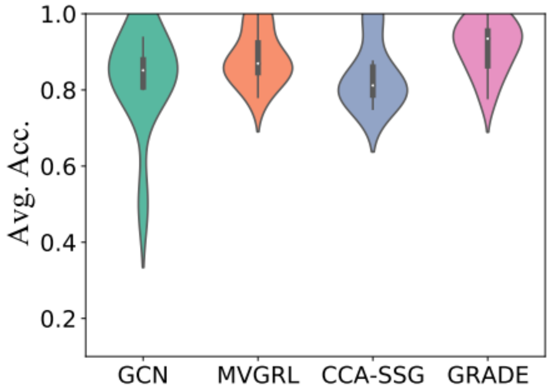
$$\mathcal{J} = \frac{1}{2N} \sum_{i=1}^{N} [\ell(\boldsymbol{h}_i, \boldsymbol{o}_i) + \ell(\boldsymbol{o}_i, \boldsymbol{h}_i)]$$

[1] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep Graph Contrastive Representation Learning. In ICML Workshop.

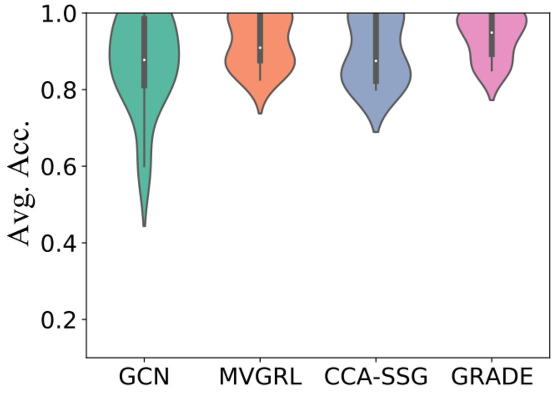
Experiments

Node Classification

		Cora		Citeseer		Photo		Computer	
		Micro-F1	*Macro-F1*	*Micro-F1*	*Macro-F1*	*Micro-F1*	*Macro-F1*	*Micro-F1*	*Macro-F1*
Supervised Split	**GCN**	82.30±0.49	76.87±0.34	65.84±0.55	59.62±0.64	93.52±0.82	**78.88**±2.01	89.14±0.75	72.61±3.05
	DGI	82.28±0.84	77.23±0.90	65.64±0.63	59.47±1.24	92.98±1.12	78.83±1.66	88.96±0.96	72.30±1.80
	GraphCL	81.78±0.67	76.01±1.07	65.16±1.02	58.72±1.37	—	—	—	—
	GRACE	82.32±0.45	76.78±0.87	64.16±2.07	59.73±1.94	93.12±0.40	78.60±3.12	88.22±1.04	71.74±3.05
	MVGRL	83.22±1.02	77.84±1.35	66.26±0.72	60.30±0.95	94.10±0.31	78.36±2.22	—	—
	CCA-SSG	82.70±0.86	77.35±1.06	65.96±1.36	58.81±1.67	94.36±0.25	79.34±3.42	89.22±0.95	73.82±1.80
	GRADE	**83.40**±0.80	**78.54**±1.15	**67.14**±1.07	**61.04**±2.07	**94.72**±0.30	78.86±2.77	**89.42**±0.53	**74.71**±1.30
Semi-supervised Split	**GCN**	74.18±0.40	69.84±0.56	53.80±0.94	50.15±0.69	91.04±0.65	65.47±1.20	78.58±0.93	61.80±1.43
	DGI	75.92±0.86	70.04±0.53	54.52±1.44	51.92±1.23	90.78±0.78	66.27±0.76	79.00±0.80	62.00±1.70
	GraphCL	75.68±2.84	69.86±2.41	54.06±1.93	51.75±1.78	—	—	—	—
	GRACE	75.12±1.41	69.66±1.29	53.56±3.42	49.83±1.74	91.12±0.31	65.07±1.28	79.10±1.79	61.76±1.97
	MVGRL	76.44±1.17	70.52±1.63	56.84±1.26	53.79±1.25	92.01±0.87	66.16±2.13	—	—
	CCA-SSG	75.74±1.96	71.70±1.59	57.90±1.82	54.70±1.54	91.68±0.50	**67.08**±1.08	82.20±0.47	65.04±1.16
	GRADE	**77.20**±0.94	**73.37**±1.27	**59.44**±0.78	**56.47**±0.64	**92.04**±0.30	66.62±2.27	**82.50**±1.04	**67.50**±1.80



(a) Tail nodes



(b) Head nodes

GRADE outperforms all baselines in most cases regardless of tail nodes or head nodes.

Fairness Analysis

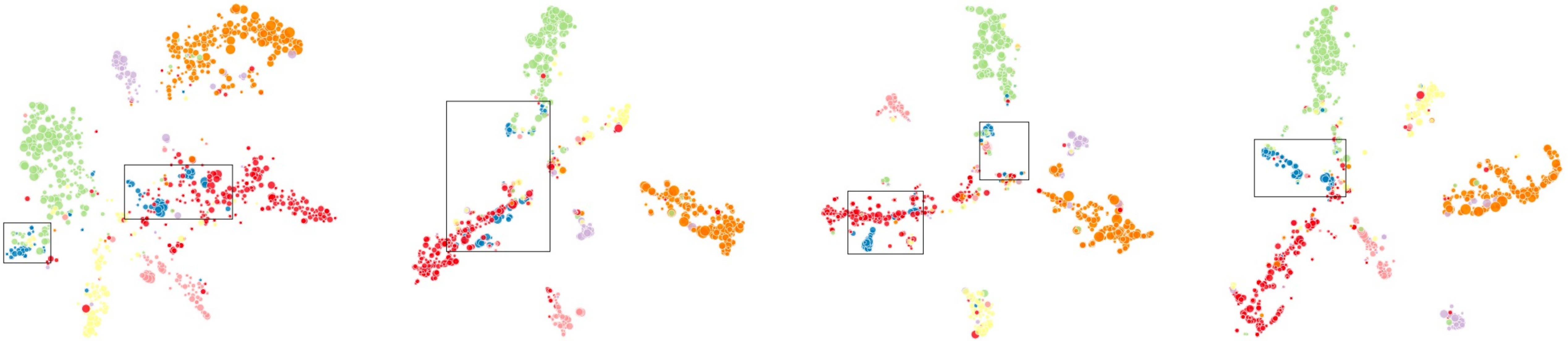
- We define the **group mean** as the mean of degree-specific average accuracy
- The **bias** is the variance.

$$\text{Avg. Acc.}(k) = \mathbb{E}[\{\text{Acc}(v_i), \forall \text{ node } v_i \text{ such that } d_i = k\}],$$
$$G.Mean = \mathbb{E}[\{\text{Avg. Acc.}(k), \forall \text{ node degree } k\}], Bias = \text{Var}(\{\text{Avg. Acc.}(k), \forall \text{ node degree } k\})$$

	Cora		Citeseer		Photo		Computer	
	G. Mean↑	*Bias*↓	*G. Mean*↑	*Bias*↓	*G. Mean*↑	*Bias*↓	*G. Mean*↑	*Bias*↓
GCN	86.04	1.70	84.00	1.85	97.41	0.28	96.30	0.50
DGI	89.26	0.67	84.79	1.71	98.23	0.27	96.94	0.45
GraphCL	90.80	0.59	84.13	1.80	—	—	—	—
GRACE	89.91	0.70	85.44	1.67	98.28	0.23	96.92	0.47
MVGRL	91.01	0.54	83.86	1.83	98.39	0.27	—	—
CCA-SSG	90.86	0.63	84.35	1.73	98.44	0.24	97.17	0.39
GRADE	**92.87**	**0.48**	**85.88**	**1.52**	**98.52**	**0.20**	**97.42**	**0.35**

GRADE reduces the bias across all datasets and maintain the highest group mean.

Visualization



(a) GCN (b) MVGRL (c) CCA-SSG (d) GRADE

GRADE pulls same-community node representations more concentrated.

Conclusions

01 New insights for structural fairness

We are the first to discover that GCL methods exhibit more structural fairness than GCN. This discovery inspires a new path for alleviating structural unfairness based on contrastive learning.

02 Deeper understanding for graph contrastive learning

We theoretically validate the reason for structural fairness in GCL is that it stimulates intra-community concentration.

03 A novel framework

We propose a method GRADE to further improve the structural fairness by enriching the neighborhood of tail nodes while purifying neighbors of head nodes.

Thanks for listening!

E-mail: wangruijia@bupt.edu.cn

code & data: http://shichuan.org/