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
{wangruijia, xiaowang, shichuan}@bupt.edu.cn, songle@biomap.com

Abstract

Recent studies show that graph convolutional network (GCN) often performs worse for low-degree nodes, exhibiting the so-called structural unfairness for graphs with long-tailed degree distributions prevalent in the real world. Graph contrastive learning (GCL), which marries the power of GCN and contrastive learning, has emerged as a promising self-supervised approach for learning node representations. How does GCL behave in terms of structural fairness? Surprisingly, we find that representations obtained by GCL methods are already fairer to degree bias than those learned by GCN. We theoretically show that this fairness stems from intra-community concentration and inter-community scatter properties of GCL, resulting in a much clear community structure to drive low-degree nodes away from the community boundary. Based on our theoretical analysis, we further devise a novel graph augmentation method, called GRAph contrastive learning for DEgree bias (GRADE), which applies different strategies to low- and high-degree nodes. Extensive experiments on various benchmarks and evaluation protocols validate the effectiveness of the proposed method.

1 Introduction

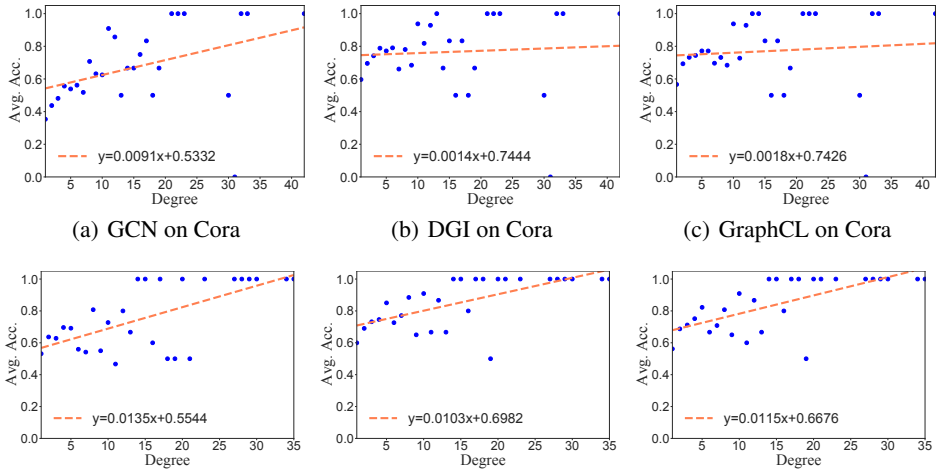
Despite their strong expressive power in graph representation learning, recent studies reveal that the performance of vanilla graph convolutional network (GCN) [17] exhibits a structural unfairness [25, 16], which is primarily beneficial to high-degree nodes (head nodes) but biased against low-degree nodes (tail nodes). Such a performance disparity is alarming and causes a performance bottleneck, given that node degrees of real-world graphs often follow a long-tailed power-law distribution [2].

Graph contrastive learning (GCL) [30, 24, 22, 38] has been a promising paradigm in graph domain, which integrates the power of GCN and contrastive leaning [13, 5, 10]. In a nutshell, these methods typically construct multiple views via stochastic augmentations of the input, then optimize the GCN encoder by contrasting positive samples against negative ones. Inheriting advantages of contrastive learning, GCL relieves graph representation learning from human annotations, and displays state-of-the-art performance in a variety of tasks [35, 12, 39, 36].

Will GCL present the same structure unfairness as GCN? For this purpose, we conduct experiments to surprisingly find out that GCL methods are better at maintaining structural fairness where a smaller performance gap exists between tail nodes and head nodes than that of GCN. This finding suggests that GCL has the potential to mitigate structural unfairness. Based on this finding, a natural and fundamental question arises: *why is graph contrastive learning fairer to degree bias?* A well-informed answer can yield profound insights into solutions to this important problem, and deepen our understanding of the mechanism of GCL.

*Corresponding authors.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



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

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

Figure 1: Visualization for the fairness of models to degree bias. Blue scatters refer to the average accuracy (Avg. Acc.) of a specific degree group. Orange dotted lines are regression lines of blue scatters in each figure. For clarity, we annotate analytical expressions of regression lines. The higher the average accuracy of tail nodes and the smaller the slope of the regression line indicate that the model is fairer to degree bias.

Intuitively, graph augmentation provides an opportunity for tail nodes to generate more within-community edges, making their representations closer to those with the same community via the contrastive framework. These refined representations drive tail nodes away from the community boundary. Theoretically, we prove that node representations learned by GCL conform to a clearer community structure by Intra-community Concentration Theorem 1 and Inter-community Scatter Theorem 2. These theorems are relevant to two crucial components in GCL. One is the alignment of positive pairs, which is exactly the optimization objective. The other is the pre-defined graph augmentation, determining the concentration of augmented representations. Based on the analysis, we establish the relation between graph augmentation and representation concentration, implying that a well-designed graph augmentation can promote structural fairness by concentrating augmented representations.

To take a step further, we propose a GRAph contrastive learning for DEgree bias (GRADE) based on a novel graph augmentation. To make augmented representations more concentrated within communities, GRADE enlarges limited neighbors of tail nodes to contain more nodes within the same community, where the ego network of the tail node interpolates with that of the sampled similar node. As for head nodes, GRADE purifies their ego networks by removing neighbors from different communities. Extensive experiments on various benchmark datasets and several evaluation protocols validate the effectiveness of GRADE.

In summary, our contributions are four-fold:

- We are the first to discover that GCL methods exhibit more structural fairness than GCN, which has a smaller performance disparity between tail nodes and head nodes. This discovery inspires a new path for alleviating structural unfairness based on contrastive learning.
- We theoretically validate the reason for structural fairness in GCL is that GCL stimulates intra-community concentration and inter-community scatter. Therefore, tail nodes are farther away from the community boundary for better classification.
- Based on theoretical insights, we propose a method GRADE to further improve structural fairness by enriching the neighborhood of tail nodes while purifying neighbors of head nodes.
- Comprehensive experiments demonstrate that our GRADE outperforms baselines on multiple benchmark datasets and enhances the fairness to degree bias.

2 Exploring the Behavior of Graph Contrastive Learning on Degree Bias

Real-world graphs in many domains follow a long-tailed distribution in node degrees, i.e., a significant fraction of nodes are tail nodes with small degrees. It is well known that GCN often performs worse accuracy for tail nodes. Here we aim to study how GCL behaves under degree bias.

Experimental Setup We take two representative GCL algorithms DGI [30] and GraphCL [35] as examples to analyze the performance disparity under degree bias. Specifically, we train DGI and GraphCL on four benchmarks Cora [17], Citeseer [17], Photo [23] and Computer [23], and leverage early stopping based on the training loss. To compare with GCN, our linear evaluation protocol deploys the semi-supervised split [17], where 20 labeled nodes per class form training set and test set composes of randomly sampled 1000 nodes with degrees less than 50. GCN follows the standard training paradigm [17] with the above train-test split. Further implementation details and chosen hyper-parameters are deferred to Appendix A.

Results To illustrate the fairness to degree bias, we group nodes of the same degree and calculate the average accuracy of these degree groups separately shown in Figure 1. To further visualize how the model balances the performance between tail nodes and head nodes, we fit these scatters with linear regression. If the slope of the regression line is flat, the model is fair to degree bias. More results on Photo and Computer datasets can be seen in Appendix A. From the figure, we can find that the average accuracy for tail nodes of GCL methods DGI and GraphCL is higher than that of GCN, and the slope of the regression line is also smaller. These interesting observations suggest that label-independent GCL methods are actually fairer than GCN under degree bias.

3 Analysis on the Structural Fairness of Graph Contrastive Learning

Based on the above observations, a natural and fundamental question arises: where does this structural fairness stem from? We first define some preliminary notations, then provide a theoretical analysis to explain this question.

3.1 Preliminary Notations

Let $G = (\mathcal{V}, \mathcal{E}, X)$ be a graph, where $\mathcal{V}$ is the set of N nodes $\{v_1, v_2, \cdots, v_N\}$, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{N \times B}$ represents the node feature matrix and $\boldsymbol{x}_i$ is the feature vector of node v_i. The edges can be represented by an adjacency matrix $A \in \{0,1\}^{N \times N}$, where $A_{ij} = 1$ iff $(v_i, v_j) \in \mathcal{E}$. Given unlabeled training nodes, each node belongs to one of K latent communities $C_1, C_2, \cdots, C_K$. Assuming the augmentation set $\mathcal{T}$ consisting of all transformations on topology, the set of potential positive samples generated from ego network $\mathcal{G}_i$ of node v_i is denoted as $\mathcal{T}(\mathcal{G}_i)$. The goal of GCL is to learn a GCN encoder f such that positive pairs are closely aligned while negative pairs are far apart. Here we focus on topological augmentation and single-layer GCN,

$$f(\mathcal{G}_i) = ReLU(\tilde{L}_i X W), \tag{1}$$

where $\tilde{L}_i$ is the i-row of transition matrix $\tilde{L} = \tilde{D}^{-1}\tilde{A}$, $\tilde{A} = A + I$ is self-looped adjacency matrix and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is degree matrix. We consider a community indicator F_f

$$F_f(\mathcal{G}_i) = \arg\min_{k \in [K]} \| f(\mathcal{G}_i) - \mu_k \|, \tag{2}$$

where $\mu_k = \mathbb{E}_{v_i \in C_k} \mathbb{E}_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i)} [f(\hat{\mathcal{G}}_i)]$ is the community center, and $\| \cdot \|$ stands for l_2-norm. To quantify the performance of F_f, the error can be formulated as

$$\mathrm{Err}(F_f) = \sum_{k=1}^{K} \mathbb{P}[F_f(\mathcal{G}_i) \neq k, \forall v_i \in C_k]. \tag{3}$$

With the above definitions, 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 the set of nodes with ε-close representations among graph augmentations.

3.2 Theoretical Analysis

We assume the nonlinear transformation has M-Lipschitz continuity, i.e., $\| f(\mathcal{G}_i) - f(\mathcal{G}_j) \| = \| ReLU(\tilde{L}_i X W) - ReLU(\tilde{L}_j X W) \| \leq M \| \tilde{L}_i X - \tilde{L}_j X \|$, and 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 $\|\tilde{L}_i X - \hat{L}_i X\|^2 \leq \beta m$, where $\hat{L}$ is the augmented transition.

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.*

This theorem builds a relation between the intra-community concentration of final representations and the alignment of positive pairs in S_ε. Specifically, intra-commmunity concentration requires smaller ε. By decreasing the distance between positive pairs, GCL fits the requirement.

Next, we demonstrate that GCL also maintains the property of inter-community scatter for community assignment. For a given augmentation set $\mathcal{T}$, we first define the augmentation distance between two nodes as the minimum distance between their pre-transformation representations,

$$d_{\mathcal{T}}(v_i, v_j) = \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \|\hat{L}_i X - \hat{L}_j X\| = \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \|(\frac{\hat{A}_i}{\hat{d}_i} - \frac{\hat{A}_j}{\hat{d}_j})X\|, \quad (4)$$

where $\hat{A}_i$ is the i-row of augmented adjacency matrix $\hat{A}$, and $\hat{d}_i$ is the augmented node degree. Based on the augmentation distance, we further 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. $\sup_{v_i, v_j \in C_k^0} d_{\mathcal{T}}(v_i, v_j) \leq \gamma(\frac{B}{\hat{d}^k_{\min}})^{\frac{1}{2}}$ *where* $\gamma \in (0, 1]$,

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

Larger α and smaller $\gamma(B/\hat{d}^k_{\min})^{\frac{1}{2}}$ indicate pre-transformation representations are more concentrated. We assume that the representation is normalized by $\|f(\mathcal{G}_i)\| = r$ and let $p_k = \mathbb{P}[v_i \in C_k]$. Then we simultaneously bound the inter-community distance and the error of the community indicator.

Lemma 1 *For a $(\alpha, \gamma, \hat{d})$-augmentation with subset C_k^0 of each community C_k, if nodes belonging to $(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon$ can be correctly assigned by the community indicator F_f, then the error of all nodes can be bounded by $(1-\alpha) + R_\varepsilon$, where $R_\varepsilon = \mathbb{P}[\overline{S_\varepsilon}]$ is the proportion of complement.*

The above lemma presents a sufficient condition to guarantee the performance of the community indicator. Then we need to explore when nodes in $(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon$ can be correctly assigned by F_f.

Lemma 2 *For a $(\alpha, \gamma, \hat{d})$-augmentation and each $\ell \in [K]$, if*

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

holds for all $k \neq \ell$, then every node $v_i \in C_\ell^0 \cap S_\varepsilon$ can be correctly assigned by the community indicator F_f, where $\rho_\ell(\alpha, \delta, \varepsilon) = 2(1-\alpha) + \frac{2R_\varepsilon}{p_\ell} + \alpha(\frac{M\gamma\sqrt{B}}{r\sqrt{\hat{d}^\ell_{\min}}} + \frac{2\varepsilon}{r})$ and $\Delta_\mu = 1 - \min_{k \in [K]} \|\mu_k\|^2/r^2$.

Combining Lemma 1 and 2, we can obtain the Inter-community Scatter Theorem as follows.

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

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

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* $\Delta_\mu = 1 - \min_{k\in[K]} \|\mu_k\|^2/r^2$.

For the better assignment, the RHS of Equation 27 should approach r^2, implying that smaller $\rho_{\max}(\alpha, \gamma, \hat{d}, \varepsilon)$ is required. We further bound R_ε by the alignment objective of contrastive loss.

Theorem 3 *The term* R_ε *is upper bounded by*

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

We direct the readers to Appendix B for proof of all the above lemmas and theorems. The combination of Theorem 2 and Theorem 3 indicates that the inter-community distance and the error of community indicator are controlled by two key factors. 1) The alignment of positive pairs. Good alignment enables small $\mathbb{E}_{\hat{\mathcal{G}}_i^1, \hat{\mathcal{G}}_i^2 \in \mathcal{T}(\mathcal{G}_i)} \|f(\hat{\mathcal{G}}_i^1) - f(\hat{\mathcal{G}}_i^2)\|$, resulting in small R_ε. 2) The concentration of augmented representations, where sharper concentration implies larger α in Definition 1. Small R_ε and large α directly decrease the error bound of community indicator, and provide small $\rho_{\max}(\alpha, \gamma, \hat{d}, \varepsilon)$ for inter-community scatter. It is worth mentioning that the first factor is the contrastive objective in GCL, reflecting the reason for structural fairness in the GCL framework. While the second factor depends on the graph augmentation. Thus, we are motivated to propose a graph augmentation designed for further concentrating augmented representations.

4 GRADE Methodology

In this section, we present our novel GRADE framework dedicated to degree bias in detail, starting with the special graph augmentation, followed by the objective of GCL.

4.1 Graph Augmentation

We generate two augmentations $\hat{G}_1$ and $\hat{G}_2$ by simultaneously corrupting the original feature and topology to construct diverse contexts [36, 26]. We denote node representations in these two augmentations as $H = f(\hat{G}_1)$ and $O = f(\hat{G}_2)$.

Topology Augmentation To obtain more concentrated augmented representations, we aim to increase intra-community edges while decreasing inter-community edges. Due to the different structural properties of tail nodes and head nodes, we design different topology augmentation strategies for them shown in Figure 2. In order to expand the neighborhood of tail nodes to include more same-community nodes, we interpolate the ego network of the anchor tail node v_{tail} with that of a sampled similar node v_{sample}. To prevent injecting many different-community neighbors, we regulate the interpolation ratio depending on the similarity between v_{tail} and v_{sample}. For head nodes, we purify their neighborhood by similarity-based sampling to remove inter-community edges.

Formally, we build the similarity matrix S between node pairs based on cosine similarity of their representations, $S_{ij} = \mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)$ for $i \neq j$ and $S_{ii} = 0$ otherwise. For any tail node v_{tail}, we sample a node v_{sample} from the multimodal distribution $\mathrm{Multi}(\boldsymbol{s}_{tail})$, where $\boldsymbol{s}_{tail}$ is the row vector of S corresponding to v_{tail}. Then we construct a new similarity-aware neighborhood for v_{tail} by interpolating between neighbor distributions of v_{tail} and v_{target}. Here, the neighbor distribution for node v is defined as $p(u|v) = 1/|\mathcal{N}(v)|$ if $u \in \mathcal{N}(v)$ and $p(u|v) = 0$ otherwise. To avoid the detrimental connectivity, the similarity $\mathrm{sim}(\boldsymbol{h}_{tail}, \boldsymbol{h}_{sample})$ is used as the interpolation ratio ϕ,

$$p_{sample}(u|v_{tail}) = \phi p(u|v_{tail}) + (1-\phi) p(u|v_{sample}). \tag{7}$$

The interpolation ratio ϕ decreases as the similarity between v_{tail} and v_{sample} decreases, and we guarantee ϕ to be at least 0.5 to preserve the original neighborhood. Given the neighbor distribution $p_{sample}(u|v_{tail})$, we sample neighbors from it without replacement. The number of neighbors is sampled from degree distribution except tail nodes to keep degree statistics.

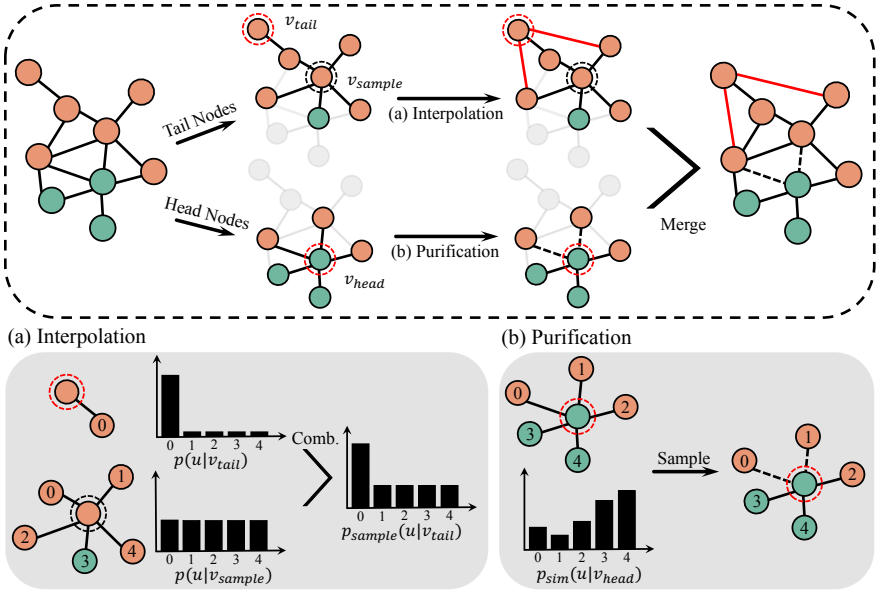


Figure 2: Topology augmentation in GRADE. Different augmentation strategies are designed for tail nodes and head nodes. Tail nodes obtain more intra-community edges by interpolation, while head nodes remove inter-community edges via purification.

For each head node v_{head}, we define the similarity distribution for purification. Specifically, the similarity distribution for node v is $p_{sim}(u|v) = \mathrm{sim}(\boldsymbol{h}_u, \boldsymbol{h}_v)$ if $u \in \mathcal{N}(v)$ and $p(u|v) = 0$ otherwise. Based on the similarity distribution $p_{sim}(u|v_{head})$, we sample $d_{head}(1 - p_{edr})$ neighbors without replacement, where p_{edr} is the edge drop rate. Through this sampling, edges of dissimilar nodes tend to be removed, thereby retaining effective neighborhood information.

Feature Augmentation We randomly sample a mask vector $\boldsymbol{m} \in \{0, 1\}^B$ to hide a fraction of dimensions in node feature. Each element in mask $\boldsymbol{m}$ is sampled from a Bernoulli distribution $\mathrm{Ber}(1 - p_{fdr})$, where the hyperparameter p_{fdr} is the feature drop rate. Thus, the augmented node feature $\hat{X}$ is computed by

$$\hat{X} = [\boldsymbol{x}_1 \circ \boldsymbol{m}, \boldsymbol{x}_2 \circ \boldsymbol{m}, \cdots, \boldsymbol{x}_N \circ \boldsymbol{m}]. \tag{8}$$

In our implementation, we set a threshold ζ to distinguish tail nodes and head nodes. The same hyperparameters p_{fdr} and p_{edr} are used to generate augmentations $\hat{G}_1$ and $\hat{G}_2$.

4.2 Optimization Objective

We employ a contrastive objective [38] on obtained node representations of two graph augmentations. For node v_i, node representations $\boldsymbol{h}_i$ and $\boldsymbol{o}_i$ from different graph augmentations form the positive pair, and node representations of other nodes in two graph augmentations are regarded as negative pairs. Therefore, we define the pairwise objective for each positive pair $(\boldsymbol{h}_i, \boldsymbol{o}_i)$ as

$$\ell(\boldsymbol{h}_i, \boldsymbol{o}_i) = \log \frac{e^{\theta(\boldsymbol{h}_i, \boldsymbol{o}_i)/\tau}}{e^{\theta(\boldsymbol{h}_i, \boldsymbol{o}_i)/\tau} + \sum_{k \neq i} e^{\theta(\boldsymbol{h}_i, \boldsymbol{o}_k)/\tau} + \sum_{k \neq i} e^{\theta(\boldsymbol{h}_i, \boldsymbol{h}_k)/\tau}}. \tag{9}$$

where τ is a temperature parameter. The critic $\theta(\boldsymbol{h}, \boldsymbol{o})$ is defined as $\mathrm{sim}(g(\boldsymbol{h}), g(\boldsymbol{o}))$, where the projection g is a two-layer multilayer perceptron (MLP) to enhance the expression power [5]. The overall objective to be maximized is the average of all positive pairs,

$$\mathcal{J} = \frac{1}{2N} \sum_{i=1}^{N} \left[\ell(\boldsymbol{h}_i, \boldsymbol{o}_i) + \ell(\boldsymbol{o}_i, \boldsymbol{h}_i) \right]. \tag{10}$$

Table 1: Quantitative results (%) on node classification. (bold: best; em dash: out-of-memory)

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

5 Experiments

Datasets For a comprehensive comparison, we use four real-world datasets to evaluate the performance of node classification and the fairness to degree bias. Specifically, we choose two categories of datasets: 1) citation networks including Cora [17] and Citeseer [17], 2) social networks Photo [23] and Computer [23] from Amazon. The statistics of these datasets are summarized in Appendix C.

Evaluation Protocol We compare GRADE with state-of-the-art GCL models DGI [30], GraphCL [35], GRACE [38], MVGRL [12] and CCA-SSG [36], and semi-supervised baseline GCN [17] for reference. For GCL models, we follow the linear evaluation scheme introduced in [30], where each model is firstly trained in an unsupervised manner and node representations are subsequently fed into a simple logistic regression classifier. We adopt two universally accepted splits for full evaluation: 1) semi-supervised split [30, 35] that 20 labeled nodes per class are for training and 1000 nodes are for testing, 2) supervised split [38, 36] that 1000 nodes are for testing and the rest of nodes form the training set. It is worth noting that 1000 nodes in the test set are randomly sampled with degrees less than 50 to provide an appropriate degree range for analysis. GCN is trained by the original paradigm [17] with the above train-test split. We refer readers of interest to Appendix C on details of experiments, including implementation and hyperparameters.

5.1 Main Results and Analysis

Node Classification We train each model for 10 independent trials with different seeds, and report mean and standard deviation results in Table 1. We observe that the proposed GRADE outperforms all baselines in most cases. The improvement of GRADE is more pronounced on Cora and Citeseer datasets, where average node degrees are around 3 and a large number of tail nodes exist. Additionally, we find that GCL models tend to outperform GCN in the semi-supervised split, suggesting that GCN may benefit more from end-to-end training with more supervision.

To verify that GRADE improves the classification performance of tail nodes and also retains the performance of head nodes, we divide test nodes of Cora into tail nodes and head nodes based on the threshold ζ. We draw average accuracy w.r.t. degree of GRADE and competitive baselines as violin plots in Figure 3. The subsequent experiments default to the supervised split if not specified. As expected, GRADE achieves obvious performance gain regardless of tail nodes or head nodes.

Fairness Analysis In order to quantitatively analyze the fairness to degree bias, we define the group mean as the mean of degree-specific average accuracy while the bias is the variance. Mathematically,

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

Table 2: Quantitative results (%) on fairness analysis.

	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**

Based on these metrics, evaluation results are shown in Table 2. It can be seen that GRADE reduces the bias across all datasets and maintain the highest group mean. Moreover, graph contrasting learning models have a smaller bias compared to GCN, conforming to our previous study in Section 2.

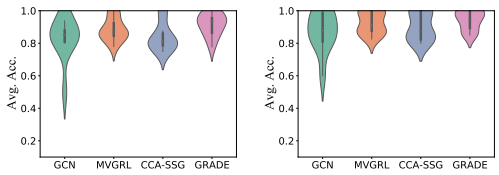
Visualization To demonstrate that GRADE pulls same-community node representations more concentrated, we visualize node representations of GRADE and competitive baselines on the Cora dataset in Figure 5. Particularly, we zoom into one specific community colored blue. Graph contrastive learning baselines always depict more crisp boundaries than GCN, where blue nodes are still scattered in space. In GRADE, they are finally clustered together, illustrating that the proposed augmentation design exerts a vital part.

5.2 Ablation Study and Hyperparameter Sensitivity

Ablation Study of Sampling Recall that GRADE augments tail nodes by sampling nodes for interpolation, and also sample edges to remove in the purification of head nodes. We alter the augmentation by fixing to the most similar node (without random interpolation) or top-$d_{head}(1-p_{edr})$ similar nodes (without random deleting) to validate the effectiveness of these sampling processes. Results of the ablation study on Cora and Citeseer datasets are reported in Table 3. We can observe that GRADE is consistently better than the remaining variants. Such a phenomenon implies that reasonable randomness provides more diverse node contexts to contrast out essential features.

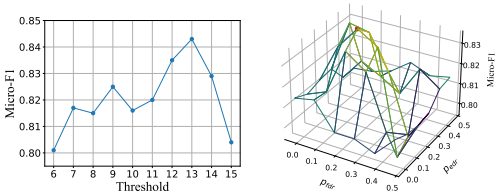
Effect of Threshold We investigate the impact of threshold ζ used to split tail nodes and head nodes on classification performance. Figure 4 (a) shows the test Micro-F1 w.r.t. different ζ on Cora dataset. The performance benefits from an applicable selection of ζ. When ζ is too small, the neighbor sparsity of tail nodes cannot be mitigated; if it is too large, noise is injected.

Effect of Drop Rate We perform sensitivity analysis on feature drop rate d_{fdr} and edge drop rate d_{edr} which control the generation of graph augmentations. We vary these hyperparameters from 0 to 0.5 in node classification on the Cora dataset. The results are shown in Figure 4 (b). It can be observed that the performance is relatively poor when the hyperparameter d_{fdr} is too large and d_{edr} is too small. We infer that if the d_{fdr} is too large, the original graph is heavily undermined to contain useful information; and if the d_{edr} is too small, the large number of neighbors makes the perturbation of different augmentations too insignificant to achieve the purpose of contrast.



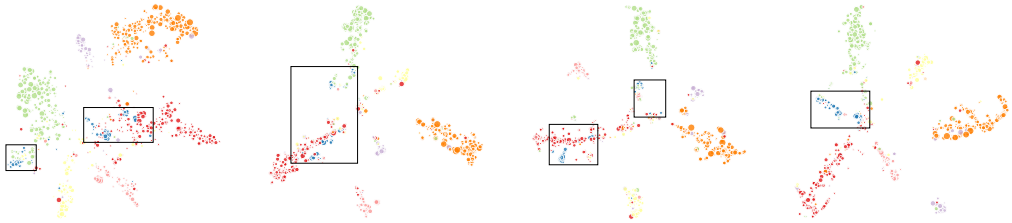
(a) Tail nodes (b) Head nodes

Figure 3: Violin plots of the average accuracy w.r.t. node degree for (a) tail nodes and (b) head nodes on the Cora dataset. The box inside the violin indicates 25-75 percentiles, and the median is shown by a white scatter.



(a) Threshold (b) Drop rate

Figure 4: The hyperparameter sensitivity of GRADE with varying (a) threshold and (b) drop rate on Cora dataset. In (b), the lighter color represents better performance, and the red scatter highlights the peak.



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

Figure 5: Visualization of node representations learned by competitive baselines and GRADE on Cora dataset. Color denotes the community of nodes and size represents the node degree. Black boxes highlight one community colored by blue as an example.

Table 3: Ablation study (%) on the sampling of GRADE. (w/o RI: without random interpolation; w/o RD: without random deleting)

	Cora				**Citeseer**			
	Micro-F1↑	*Macro-F1*↑	*G. Mean*↑	*Bias*↓	*Micro-F1*↑	*Macro-F1*↑	*G. Mean*↑	*Bias*↓
w/o RI	83.00	77.97	90.86	0.55	66.10	61.02	84.96	1.80
w/o RD	82.30	76.50	90.54	0.60	65.90	58.83	84.21	1.84
w/o RI+RD	82.00	75.82	89.97	0.63	65.80	59.29	83.00	1.95
GRADE	**84.30**	**80.66**	**92.87**	**0.48**	**68.80**	**64.21**	**85.88**	**1.52**

6 Related Work

Graph Neural Networks Graph neural networks (GNNs) [18, 32, 4, 31] can be generally divided into spectral methods and spatial methods. Specifically, spectral methods learn node representations based on graph spectral theory. [3] first proposes a spectral graph-based extension of convolutional networks, and GCN [17] simplifies ChebNet [6] by the first-order approximation. Spatial methods directly define graph convolution in the spatial domain. GraphSAGE [11] learns aggregators by sampling and aggregating neighbors, and GAT [29] assigns different edge weights during aggregation. We refer readers to recent surveys [34, 37] for a more comprehensive review. On the other hand, some literature [25, 33, 21] shows that there is a structural unfairness between tail nodes and head nodes in GNNs. Existing explorations [20, 16] focus on supervised settings, ignoring the great potential of self-supervised learning on this problem.

Graph Contrastive Learning Being popular in self-supervised visual representation learning [28, 7, 27, 13, 5, 10, 15], contrastive learning obtains discriminative representations by contrasting positive and negative samples. Inspired by the local-global mutual information maximization viewpoint [14], DGI [30] and InfoGraph [24] first marry the power of GNNs and contrastive learning. Following them, MVGRL [12] introduces the node diffusion to the graph contrastive framework. GRACE [38], GCA [39] and GraphCL [35] learn node representations by treating other nodes as negative samples, while BGRL [26] proposes a negative-sample-free model. CCA-SSG [36] optimizes a feature-level objective other than instance-level discrimination. There are several surveys [1, 19] summarizing recent advances in graph contrastive learning. Despite their remarkable achievements, there is no graph contrastive learning targeting the fairness of degree bias.

7 Conclusion

In this paper, we bring to light the prospect of GCL to alleviate structural unfairness for node representation learning. We discover that node representations obtained by GCL methods are fairer to degree bias than those learned by GCN, and explore the underlying cause of this phenomenon. Based on our theoretical analysis, we further propose a novel GCL model targeting degree bias.

Limitations and Broader Impact A limitation of GRADE is its heuristic design, therefore an interesting direction for future work is to extend GRADE into learnable graph augmentation. Our work investigates the structural fairness in GCL for the first time and points out the great potential of GCL for this problem. Considering that most real-world graphs follow the long-tail distribution, the studied problem is practical and important. Additionally, our work deepens the understanding of the learning mechanism of GCL, and may inspire more future research on structural fairness.

Acknowledgments and Disclosure of Funding

We thank anonymous reviewers for their time and effort in reviewing this paper. This work is supported in part by the National Natural Science Foundation of China (No. U20B2045, 62192784, 62172052, 62002029, U1936014).

References

[1] Self-supervised learning on graphs: Contrastive, generative, or predictive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and deep locally connected networks on graphs. In *ICLR*, 2014.

[4] Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: Fast learning with graph convolutional networks via importance sampling. In *ICLR*, 2018.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[6] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*, 2016.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[8] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 1960.

[9] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *NeurIPS*, 2020.

[11] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.

[12] Kaveh Hassani and Amir Hosein Khas Ahmadi. Contrastive multi-view representation learning on graphs. In *ICML*, 2020.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[14] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2018.

[15] Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning. *arXiv e-prints*, 2021.

[16] Jian Kang, Yan Zhu, Yinglong Xia, Jiebo Luo, and Hanghang Tong. Rawlsgcn: Towards rawlsian difference principle on graph convolutional network. In *WWW*, 2022.

[17] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[18] Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *ICLR*, 2018.

[19] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and Philip Yu. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[20] Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. Tail-gnn: Tail-node graph neural networks. In *KDD*, 2021.

[21] Zemin Liu, Wentao Zhang, Yuan Fang, Xinming Zhang, and Steven CH Hoi. Towards locality-aware meta-learning of tail node embeddings on networks. In *CIKM*, 2020.

[22] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *KDD*, 2020.

[23] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. In *NeurIPS Workshop on Relational Representation Learning*, 2018.

[24] Fan-Yun Sun, Jordon Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2020.

[25] Xianfeng Tang, Huaxiu Yao, Yiwei Sun, Yiqi Wang, Jiliang Tang, Charu Aggarwal, Prasenjit Mitra, and Suhang Wang. Investigating and mitigating degree-related biases in graph convoltuional networks. In *CIKM*, 2020.

[26] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Rémi Munos, Petar Veličković, and Michal Valko. Bootstrapped representation learning on graphs. In *ICLR Workshop on Geometrical and Topological Representation Learning*, 2021.

[27] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.

[28] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, 2018.

[29] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[30] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *ICLR*, 2019.

[31] Ruijia Wang, Shuai Mou, Xiao Wang, Wanpeng Xiao, Qi Ju, Chuan Shi, and Xing Xie. Graph structure estimation neural networks. In *WWW*, 2021.

[32] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *ICML*, 2019.

[33] Jun Wu, Jingrui He, and Jiejun Xu. Net: Degree-specific graph neural networks for node and graph classification. In *KDD*, 2019.

[34] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[35] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

[36] Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S Yu. From canonical correlation analysis to self-supervised graph neural networks. In *NeurIPS*, 2021.

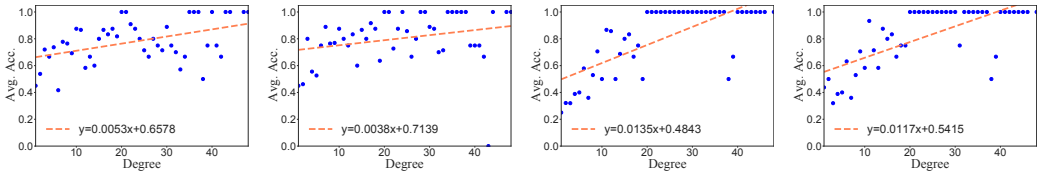
[37] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[38] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep Graph Contrastive Representation Learning. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020.

[39] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *WWW*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3 and Appendix B.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2, Section 5, Appendix A and C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 2 and Section 5.
   (b) Did you mention the license of the assets? [Yes] See Appendix C.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix C.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix C.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]



(a) GCN on Photo (b) DGI on Photo (c) GCN on Computer (d) DGI on Computer

Figure 6: More results for the fairness of models to degree bias on Photo and Computer datasets.

A Details of Section 2

A.1 Implementation Details

We choose the commonly used Cora [17] , Citeseer [17] , Photo [23] and Computer [23] for evaluation. In Cora and Citeseer datasets, nodes represent papers, edges are the citation relationship between papers, node features comprise bag-of-words vector of keywords and labels represent the research field of papers. Photo and Computer datasets are segments of the Amazon co-purchase graph, where nodes represent products, edges indicate that two products are frequently bought together, node features are bag-of-words encoded product reviews and labels are given by the product category. The statistics of these datasets are summarized in Table 4. The above datasets are public and do not contain personally identifiable information and offensive content. The URL of our datasets is https://docs.dgl.ai/api/python/dgl.data.html and the license is Apache License 2.0.

We train DGI[2] [30] and GraphCL[3] [35] on these datasets with codes provided by authors. To compare with GCN[4] [17], our linear evaluation protocol deploys the semi-supervised split [17], where 20 labeled nodes per class form training set and test set composes of randomly sampled 1000 nodes with degree less than 50. GCN follows the standard training paradigm [17] with the above train-test split. All these methods are initialized as the corresponding papers and consist of two GCN layers, where their hyperparameters are carefully searched to achieve optimal performance on the test set.

A.2 Additional Results

More comparison results between GCL methods and GCN on Photo and Computer datasets are shown in Figure 6. Please note that GraphCL has an out-of-memory issue on these datasets. From the figure, the gap between the slopes of DGI and GCN is relatively small. A reasonable hypothesis is that average node degrees of Photo and Computer datasets are much larger than those of Cora and Citeseer datasets, where the advantage of GCL to alleviate the neighborhood sparsity of tail nodes cannot be well exhibited.

B Details of Section 3

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 $\sigma^2_{f,\varepsilon} \leq \frac{1}{\kappa}\sigma^2$.*

Proof For node v_i in S_ε, we have $\|f(\mathcal{G}_i) - f(\hat{\mathcal{G}}_i)\|^2 \leq \varepsilon^2$. This implies that for nodes $v_i, v_j \in S_\varepsilon$ such that $\|\tilde{L}_i X - \tilde{L}_j X\|^2 \leq 2\beta m$, there exists a region of overlap so that $\|f(\mathcal{G}_i) - f(\mathcal{G}_j)\|^2 \leq \|f(\mathcal{G}_i) - f(\hat{\mathcal{G}}_i)\|^2 + \|f(\hat{\mathcal{G}}_i) - f(\mathcal{G}_j)\|^2 \leq 2\varepsilon^2$. That is, there are graph augmentations of v_i which are sufficiently similar to graph augmentations of v_j so that their representations should be similar, thereby driving $f(\mathcal{G}_i)$ and $f(\mathcal{G}_j)$ to be closer.

[2](MIT license) https://github.com/PetarV-/DGI
[3](MIT license) https://github.com/Shen-Lab/GraphCL
[4](MIT license) https://github.com/tkipf/pygcn

The variance of ε-close node representations in f space is

$$\sigma_{f,\varepsilon}^2 = \frac{1}{2N^2(1-R_\varepsilon)^2} \sum_{v_i \in S_\varepsilon} \sum_{v_j \in S_\varepsilon} \|f(\mathcal{G}_i) - f(\mathcal{G}_j)\|^2. \tag{11}$$

The overlap $\beta m < \|\tilde{L}_i X - \tilde{L}_j X\|^2 \leq 2\beta m$ induces a graph where we say $v_j \in \mathcal{N}(i) \forall v_j$ s.t. $\|\tilde{L}_i X - \tilde{L}_j X\|^2 \leq 2\beta m$. For $N(1-R_\varepsilon)$ samples, we can decompose the variance as

$$\begin{aligned} \sigma_{f,\varepsilon}^2 &= \frac{1}{2N^2(1-R_\varepsilon)^2} \sum_{v_i \in S_\varepsilon} \sum_{v_j \in S_\varepsilon} \|f(\mathcal{G}_i) - f(\mathcal{G}_j)\|^2 \\ &= \frac{1}{2N^2(1-R_\varepsilon)^2} \sum_{v_i \in S_\varepsilon} \left(\sum_{v_j \in \mathcal{N}(i)} \|f(\mathcal{G}_i) - f(\mathcal{G}_j)\|^2 + \sum_{v_j' \notin \mathcal{N}(i)} \|f(\mathcal{G}_i) - f(\mathcal{G}_{j'})\|^2 \right). \end{aligned} \tag{12}$$

By the smoothness of f we always have $\|f(\mathcal{G}_i) - f(\mathcal{G}_{j'})\|^2 \leq M^2 \|\tilde{L}_i X - \tilde{L}_{j'} X\|^2$. By the constraint we have that $\|f(\mathcal{G}_i) - f(\mathcal{G}_j)\|^2 \leq \frac{2\varepsilon^2 M^2}{\beta m} \|\tilde{L}_i X - \tilde{L}_j X\|^2 \forall v_j \in \mathcal{N}(i)$ and for $\eta = \frac{2\varepsilon^2 M^2}{\beta m} < 1$.

Assuming that there is a constant proportion $0 \leq \lambda \leq 1$ of nodes in the set $\mathcal{N}(i) \forall v_i \in S_\varepsilon$, thus this graph is an Erdős-Renyi graph. From Theorem 4, if $\lambda \geq \frac{c \log N}{N}$ for $c > 1$ then with high probability, there are no unconnected components in graph. Every node is reachable from any other nodes in a finite number of steps. We can then decompose nodes in the graph into adjacent ones and those which are reachable within a certain number of steps. Let the shortest path between any two nodes be at most D, then we obtain the following inequality

$$\begin{aligned} \sigma_{f,\varepsilon}^2 &= \frac{1}{2N^2(1-R_\varepsilon)^2} \sum_{v_i \in S_\varepsilon} \sum_{v_j \in S_\varepsilon} \|f(\mathcal{G}_i) - f(\mathcal{G}_j)\|^2 \\ &\leq \lambda \eta \sigma_x^2 + (1-\lambda) D \eta \sigma_x^2. \end{aligned} \tag{13}$$

From Theorem 5, we have $3 \leq D \leq 4$ with high probability. So for $\sigma_{f,\varepsilon}^2 \leq \frac{1}{\kappa} \sigma_x^2$ with $\kappa \geq 1$, we require $\varepsilon^2 \leq \frac{\beta m}{2M^2 \kappa (3-2\lambda)} \leq \frac{\beta m}{6M^2 \kappa}$. □

Theorem 4 *[8] If $p = \frac{c \log N}{N}$ where $c > 1$ with high probability, then the graph $G(N,p)$ has no unconnected components.*

Theorem 5 *[9] Let $q \geq 2$ be a fixed positive integer. For $c > 0$ and*

$$p^q N^{q-1} = \log(\frac{N^2}{c}). \tag{14}$$

Then $diam(G_{N,p}) \geq q$ with probability $e^{-\frac{c}{2}}$ and $diam(G_{N,p}) \leq q+1$ with probability $1 - e^{-\frac{c}{2}}$.

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.

Remark Since the node feature X has been mapped to surface of the unit sphere $\mathbb{S}^{B-1} = \{X_i \in \mathbb{R}^B : \|X_i\| = 1\}$, there is a natural supremum for $d_{\mathcal{T}}(v_i, v_j)$ bounded by the node degree,

$$\begin{aligned} d_{\mathcal{T}}(v_i, v_j) &= \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \|(\frac{\hat{A}_i}{\hat{d}_i} - \frac{\hat{A}_j}{\hat{d}_j})X\| \\ &\leq \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \|\frac{\hat{A}_i}{\hat{d}_i} - \frac{\hat{A}_j}{\hat{d}_j}\| \cdot \sqrt{B} \\ &= \sqrt{B} \cdot \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \|\frac{\hat{A}_i}{\hat{d}_i} - \frac{\hat{A}_j}{\hat{d}_i} + \frac{\hat{A}_j}{\hat{d}_i} - \frac{\hat{A}_j}{\hat{d}_j}\| \\ &\leq \sqrt{B} \cdot \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \frac{\|\hat{A}_i - \hat{A}_j\|}{\hat{d}_i} + \|\hat{A}_j\| \left| \frac{1}{\hat{d}_i} - \frac{1}{\hat{d}_j} \right|. \end{aligned} \tag{15}$$

Without loss of generality, we assume that $\hat{d}_i \geq \hat{d}_j$

$$\begin{aligned} d_{\mathcal{T}}(v_i, v_j) &= \sqrt{B} \cdot \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \frac{\|\hat{A}_i - \hat{A}_j\|}{\hat{d}_i} + \|\hat{A}_j\| \left(\frac{1}{\hat{d}_j} - \frac{1}{\hat{d}_i} \right) \\ &\leq \sqrt{B} \cdot \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \frac{\sqrt{\hat{d}_i + \hat{d}_j}}{\hat{d}_i} + \sqrt{\hat{d}_j} \left(\frac{1}{\hat{d}_j} - \frac{1}{\hat{d}_i} \right) \\ &= \sqrt{B} \cdot \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \frac{\sqrt{\hat{d}_i + \hat{d}_j} - \sqrt{\hat{d}_j}}{\hat{d}_i} + \frac{1}{\sqrt{\hat{d}_j}} \\ &\leq \sqrt{B} \cdot \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \frac{1}{\sqrt{\hat{d}_i}} + \frac{1}{\sqrt{\hat{d}_j}} \\ &\leq 2\sqrt{B} \cdot \min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \frac{1}{\sqrt{\hat{d}_j}}. \end{aligned} \tag{16}$$

Following this form, we define the RHS of the second condition to delineate the concentrated part.

Lemma 1 *For a $(\alpha, \gamma, \hat{d})$-augmentation with subset C_k^0 of each community C_k, if nodes belonging to $(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon$ can be correctly assigned by the community indicator F_f, then the error of all nodes can be bounded by $(1-\alpha) + R_\varepsilon$, where $R_\varepsilon = \mathbb{P}[\overline{S_\varepsilon}]$ is the proportion of complement.*

Proof Since every node $v_i \in (C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon$ can be correctly assigned by F_f, the error rate

$$\begin{aligned} \mathrm{Err}(F_f) &= \sum_{k=1}^{K} \mathbb{P}[F_f(\mathcal{G}_i) \neq k, \forall v_i \in C_k] \\ &\leq \mathbb{P}\left[\overline{(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon}\right] \\ &= \mathbb{P}\left[\overline{C_1^0 \cup \cdots \cup C_K^0} \cup \overline{S_\varepsilon}\right] \\ &\leq (1-\alpha) + \mathbb{P}[\overline{S_\varepsilon}] \\ &= (1-\alpha) + R_\varepsilon. \end{aligned} \tag{17}$$

□

Lemma 2 *For a $(\alpha, \gamma, \hat{d})$-augmentation and each $\ell \in [K]$, if*

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

holds for all $k \neq \ell$, then every node $v_i \in C_\ell^0 \cap S_\varepsilon$ can be correctly assigned by the community indicator F_f, where $\rho_\ell(\alpha, \delta, \varepsilon) = 2(1-\alpha) + \frac{2R_\varepsilon}{p_\ell} + \alpha(\frac{M\gamma\sqrt{B}}{r\sqrt{\hat{d}_{\min}^\ell}} + \frac{2\varepsilon}{r})$ and $\Delta_\mu = 1 - \min_{k \in [K]} \|\mu_k\|^2 / r^2$.

Proof To show that every node $v_i \in C_\ell^0 \cap S_\varepsilon$ can be correctly assigned by F_f, we need to prove that for all $k \neq \ell$, $\|f(\mathcal{G}_i) - \mu_\ell\| < \|f(\mathcal{G}_i) - \mu_k\|$. It is equivalent to prove

$$f(\mathcal{G}_i)^\top \mu_\ell - f(\mathcal{G}_i)^\top \mu_k - (\frac{1}{2}\|\mu_\ell\|^2 - \frac{1}{2}\|\mu_k\|^2) > 0. \tag{18}$$

Let $\tilde{f}(\mathcal{G}_i) = \mathbb{E}_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i)}[f(\hat{\mathcal{G}}_i)]$. Then $\|\tilde{f}(\mathcal{G}_i)\| = \|\mathbb{E}_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i)}[f(\hat{\mathcal{G}}_i)]\| \leq \mathbb{E}_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i)}[\|f(\hat{\mathcal{G}}_i)\|] = r$. One the one hand,

$$\begin{aligned}
f(\mathcal{G}_i)^\top \mu_\ell &= \frac{1}{p_\ell} f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j}[\tilde{f}(\mathcal{G}_j)\mathbb{I}(v_j \in C_\ell)] \\
&= \frac{1}{p_\ell} f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j}[\tilde{f}(\mathcal{G}_j)\mathbb{I}(v_j \in C_\ell \cap C_\ell^0 \cap S_\varepsilon)] + \frac{1}{p_\ell} f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j}\left[\tilde{f}(\mathcal{G}_j)\mathbb{I}\left(v_j \in C_\ell \cap \overline{C_\ell^0 \cap S_\varepsilon}\right)\right] \\
&= \frac{\mathbb{P}[C_\ell^0 \cap S_\varepsilon]}{p_\ell} f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon}[\tilde{f}(\mathcal{G}_j)] + \frac{1}{p_\ell}\mathop{\mathbb{E}}_{v_j}\left[f(\mathcal{G}_i)^\top \tilde{f}(\mathcal{G}_j) \cdot \mathbb{I}(v_j \in C_\ell \backslash C_\ell^0 \cap S_\varepsilon)\right] \\
&\geq \frac{\mathbb{P}[C_\ell^0 \cap S_\varepsilon]}{p_\ell} f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon}[\tilde{f}(\mathcal{G}_j)] - \frac{r^2}{p_\ell}\mathbb{P}[C_\ell \backslash C_\ell^0 \cap S_\varepsilon],
\end{aligned} \tag{19}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Note that

$$\mathbb{P}[C_\ell \backslash C_\ell^0 \cap S_\varepsilon] \leq \mathbb{P}[(C_\ell \backslash C_\ell^0) \cup \overline{S_\varepsilon}] = (1-\alpha)p_\ell + R_\varepsilon, \tag{20}$$

and

$$\mathbb{P}[C_\ell^0 \cap S_\varepsilon] = \mathbb{P}[C_\ell] - \mathbb{P}[C_\ell \backslash C_\ell^0 \cap S_\varepsilon] \geq p_\ell - ((1-\alpha)p_\ell + R_\varepsilon) = \alpha p_\ell - R_\varepsilon. \tag{21}$$

Plugging to Eq. (19), we have

$$\begin{aligned}
f(\mathcal{G}_i)^\top \mu_\ell &\geq \frac{\mathbb{P}[C_\ell^0 \cap S_\varepsilon]}{p_\ell} f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon}[\tilde{f}(\mathcal{G}_j)] - \frac{r^2}{p_\ell}\mathbb{P}[C_\ell \backslash C_\ell^0 \cap S_\varepsilon] \\
&\geq \left(\alpha - \frac{R_\varepsilon}{p_\ell}\right) f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon}[\tilde{f}(\mathcal{G}_j)] - r^2\left(1 - \alpha + \frac{R_\varepsilon}{p_\ell}\right).
\end{aligned} \tag{22}$$

Notice that $v_i \in C_\ell^0 \cap S_\varepsilon$. For any $v_j \in C_\ell^0 \cap S_\varepsilon$, we have $d_{\mathcal{T}}(v_i, v_j) \leq \gamma(\frac{B}{\hat{d}_{\min}^\ell})^{\frac{1}{2}}$. Let $(\hat{\mathcal{G}}_i^*, \hat{\mathcal{G}}_j^*) = \arg\min_{\hat{\mathcal{G}}_i \in \mathcal{T}(\mathcal{G}_i), \hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)} \|f(\hat{\mathcal{G}}_i) - f(\hat{\mathcal{G}}_j)\|$, thus $\|f(\hat{\mathcal{G}}_i^*) - f(\hat{\mathcal{G}}_j^*)\| \leq M\gamma(\frac{B}{\hat{d}_{\min}^\ell})^{\frac{1}{2}}$. Since $v_j \in S_\varepsilon$, for any $\hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)$, $\|f(\hat{\mathcal{G}}_j) - f(\hat{\mathcal{G}}_j^*)\| \leq \varepsilon$. Similarly, since $v_i \in S_\varepsilon$, we have $\|f(\hat{\mathcal{G}}_i) - f(\hat{\mathcal{G}}_i^*)\| \leq \varepsilon$. The first term of Eq. (22) can be bounded by

$$\begin{aligned}
f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon}[\tilde{f}(\mathcal{G}_j)] &= \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)}[f(\mathcal{G}_i)^\top f(\hat{\mathcal{G}}_j)] \\
&= \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)}[f(\mathcal{G}_i)^\top (f(\hat{\mathcal{G}}_j) - f(\mathcal{G}_i) + f(\mathcal{G}_i))] \\
&\geq r^2 + \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)}[f(\mathcal{G}_i)^\top (f(\hat{\mathcal{G}}_j) - f(\mathcal{G}_i))] \\
&= r^2 + \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon} \mathop{\mathbb{E}}_{\hat{\mathcal{G}}_j \in \mathcal{T}(\mathcal{G}_j)}[f(\mathcal{G}_i)^\top (\underbrace{f(\hat{\mathcal{G}}_j) - f(\hat{\mathcal{G}}_j^*)}_{\|\cdot\| \leq \varepsilon} + \underbrace{f(\hat{\mathcal{G}}_j^*) - f(\hat{\mathcal{G}}_i^*)}_{\|\cdot\| \leq M\gamma(\frac{B}{\hat{d}_{\min}^\ell})^{\frac{1}{2}}} \\
&\quad + \underbrace{f(\hat{\mathcal{G}}_i^*) - f(\hat{\mathcal{G}}_i)}_{\|\cdot\| \leq \varepsilon})] \\
&= r^2 - r(M\gamma(\frac{B}{\hat{d}_{\min}^\ell})^{\frac{1}{2}} + 2\varepsilon).
\end{aligned} \tag{23}$$

Therefore, Eq. (22) turns to

$$
\begin{aligned}
f(\mathcal{G}_i)^\top \mu_\ell &\geq \left(\alpha - \frac{R_\varepsilon}{p_\ell}\right) f(\mathcal{G}_i)^\top \mathop{\mathbb{E}}_{v_j \in C_\ell^0 \cap S_\varepsilon} [\tilde{f}(\mathcal{G}_j)] - r^2 \left(1 - \alpha + \frac{R_\varepsilon}{p_\ell}\right) \\
&\geq \left(\alpha - \frac{R_\varepsilon}{p_\ell}\right) (r^2 - r(M\gamma(\frac{B}{\hat{d}_{\min}^\ell})^{\frac{1}{2}} + 2\varepsilon)) - r^2 \left(1 - \alpha + \frac{R_\varepsilon}{p_\ell}\right) \\
&= r^2 \left(1 - 2(1-\alpha) - \frac{2R_\varepsilon}{p_\ell} - \left(\alpha - \frac{R_\varepsilon}{p_\ell}\right) \left(\frac{M\gamma\sqrt{B}}{r\sqrt{\hat{d}_{\min}^\ell}} + \frac{2\varepsilon}{r}\right)\right) \\
&= r^2(1 - \rho_\ell(\alpha, \gamma, \hat{d}, \varepsilon)).
\end{aligned} \tag{24}
$$

On the other hand,

$$
\begin{aligned}
f(\mathcal{G}_i)^\top \mu_k &= (f(\mathcal{G}_i) - \mu_\ell)^\top \mu_k + \mu_\ell^\top \mu_k \\
&\leq \|f(\mathcal{G}_i) - \mu_\ell\| \cdot \|\mu_k\| + \mu_\ell^\top \mu_k \\
&\leq r\sqrt{\|f(\mathcal{G}_i)\|^2 - 2f(\mathcal{G}_i)^\top \mu_\ell + \|\mu_\ell\|^2} + \mu_\ell^\top \mu_k \\
&\leq r\sqrt{2r^2 - 2f(\mathcal{G}_i)^\top \mu_\ell} + \mu_\ell^\top \mu_k \\
&\leq \sqrt{2\rho_\ell(\alpha, \gamma, \hat{d}, \varepsilon)} r^2 + \mu_\ell^\top \mu_k.
\end{aligned} \tag{25}
$$

Note that $\Delta_\mu = 1 - \min_k \|\mu_k\|^2/r^2$, the LHS of Eq. (18) is

$$
\begin{aligned}
f(\mathcal{G}_i)^\top \mu_\ell - f(\mathcal{G}_i)^\top \mu_k - (\frac{1}{2}\|\mu_\ell\|^2 - \frac{1}{2}\|\mu_k\|^2) &\geq f(\mathcal{G}_i)^\top \mu_\ell - f(\mathcal{G}_i)^\top \mu_k - \frac{1}{2} r^2 \Delta_\mu \\
&\geq r^2(1 - \rho_\ell(\alpha, \gamma, \hat{d}, \varepsilon)) - \sqrt{2\rho_\ell(\alpha, \gamma, \hat{d}, \varepsilon)} r^2 - \mu_\ell^\top \mu_k - \frac{1}{2} r^2 \Delta_\mu \\
&= r^2 \left(1 - \rho_\ell(\alpha, \gamma, \hat{d}, \varepsilon) - \sqrt{2\rho_\ell(\alpha, \gamma, \hat{d}, \varepsilon)} - \frac{1}{2}\Delta_\mu\right) - \mu_\ell^\top \mu_k > 0.
\end{aligned} \tag{26}
$$

□

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

$$
\mu_\ell^\top \mu_k < r^2(1 - \rho_{\max}(\alpha, \gamma, \hat{d}, \varepsilon) - \sqrt{2\rho_{\max}(\alpha, \gamma, \hat{d}, \varepsilon)} - \frac{\Delta_\mu}{2}) \tag{27}
$$

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 $\rho_{\max}(\alpha, \gamma, \hat{d}, \varepsilon) = 2(1-\alpha) + \max_\ell \left(\frac{2R_\varepsilon}{p_\ell} + \frac{M\alpha\gamma\sqrt{B}}{r\sqrt{\hat{d}_{\min}^\ell}}\right) + \frac{2\alpha\varepsilon}{r})$ and $\Delta_\mu = 1 - \min_{k\in[K]} \|\mu_k\|^2/r^2$.

Proof Since the augmentation $\mathcal{T}$ is $(\alpha, \gamma, \hat{d})$-augmentation, there exists a subset C_k^0 for each community C_k such that $\mathbb{P}[C_k^0] \geq \alpha p_k$ and $\sup_{v_i, v_j \in C_k^0} d_{\mathcal{T}}(v_i, v_j) \leq \gamma(\frac{B}{\hat{d}_{\min}^k})^{\frac{1}{2}}$. Since for any $\ell \neq k$, we have $\mu_\ell^\top \mu_k < r^2(1 - \rho_{\max}(\alpha, \gamma, \hat{d}, \varepsilon) - \sqrt{2\rho_{\max}(\alpha, \gamma, \hat{d}, \varepsilon)} - \frac{\Delta_\mu}{2}) \leq r^2(1 - \rho_\ell(\alpha, \gamma, \hat{d}, \varepsilon) - \sqrt{2\rho_\ell(\alpha, \gamma, \hat{d}, \varepsilon)} - \frac{\Delta_\mu}{2})$. According to Lemma 2, every node $v_i \in C_\ell^0 \cap S_\varepsilon$ can be correctly assigned by F_f. Therefore, every node $v_i \in (C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon$ can be correctly assigned by F_f. According to Lemma 1, the error rate $\mathrm{Err}(F_f) \leq (1-\alpha) + R_\varepsilon$. □

Theorem 3 *The term R_ε is upper bounded by*

$$
R_\varepsilon \leq \frac{[C(N-1, m)]^2}{\varepsilon} \mathbb{E}_{v_i} \mathbb{E}_{\hat{\mathcal{G}}_i^1, \hat{\mathcal{G}}_i^2 \in \mathcal{T}(\mathcal{G}_i)} \|f(\hat{\mathcal{G}}_i^1) - f(\hat{\mathcal{G}}_i^2)\|. \tag{28}
$$

Proof For any given node v_i, we have

$$
\sup_{\hat{\mathcal{G}}_i^1, \hat{\mathcal{G}}_i^2 \in \mathcal{T}(\mathcal{G}_i)} \|f(\hat{\mathcal{G}}_i^1) - f(\hat{\mathcal{G}}_i^2)\| \geq [C(N-1, m)]^2 \mathop{\mathbb{E}}_{\hat{\mathcal{G}}_i^1, \hat{\mathcal{G}}_i^2 \in \mathcal{T}(\mathcal{G}_i)} \|f(\hat{\mathcal{G}}_i^1) - f(\hat{\mathcal{G}}_i^2)\|. \tag{29}
$$

Table 4: Statistics of datasets.

Dataset	# Nodes	# Edges	# Features	# Classes	# Avg. Degree
Cora	2,708	10,556	1,433	7	3.89
Citeseer	3,327	9,228	3,703	6	2.77
Photo	7,650	238,163	745	8	31.13
Computer	13,752	491,722	767	10	35.75

Therefore, the following set S is a subset of S_ε,

$$S = \left\{ v_i : \mathop{\mathbb{E}}_{\hat{\mathcal{G}}_i^1, \hat{\mathcal{G}}_i^2 \in \mathcal{T}(\mathcal{G}_i)} \|f(\hat{\mathcal{G}}_i^1) - f(\hat{\mathcal{G}}_i^2)\| \leq \frac{\varepsilon}{[C(N-1,m)]^2} \right\} \subseteq S_\varepsilon. \tag{30}$$

By Markov's inequality, we have

$$\begin{aligned} R_\varepsilon = \mathbb{P}[\overline{S_\varepsilon}] &\leq \mathbb{P}[\bar{S}] \\ &\leq \frac{\mathbb{E}_{v_i} \mathbb{E}_{\hat{\mathcal{G}}_i^1, \hat{\mathcal{G}}_i^2 \in \mathcal{T}(\mathcal{G}_i)} \|f(\hat{\mathcal{G}}_i^1) - f(\hat{\mathcal{G}}_i^2)\|}{\frac{\varepsilon}{[C(N-1,m)]^2}} \\ &= \frac{[C(N-1,m)]^2}{\varepsilon} \mathbb{E}_{v_i} \mathbb{E}_{\hat{\mathcal{G}}_i^1, \hat{\mathcal{G}}_i^2 \in \mathcal{T}(\mathcal{G}_i)} \|f(\hat{\mathcal{G}}_i^1) - f(\hat{\mathcal{G}}_i^2)\|. \end{aligned} \tag{31}$$

□

C Details of Section 5

Baselines We compare GRADE with state-of-the-art GCL models DGI [30], GraphCL [35], GRACE[5] [38], MVGRL[6] [12] and CCA-SSG[7] [36] and semi-supervised GCN [17] with their original codes. For GCL models, we follow the linear evaluation scheme introduced in [30], where each model is firstly trained in an unsupervised manner and node representations are subsequently fed into a simple logistic regression classifier. We adopt two universally accepted splits for full evaluation: 1) semi-supervised split [30, 35] that 20 labeled nodes per class are for training and 1000 nodes are for testing, 2) supervised split [38, 36] that 1000 nodes are for testing and the rest of nodes form the training set. It is worth noting that 1000 nodes in the test set are randomly sampled with degrees less than 50 to provide an appropriate degree range for analysis. GCN is trained by the original paradigm [17] with the above train-test split. All these methods are initialized as the corresponding papers and consist of two GCN layers, where their hyperparameters are carefully searched to achieve optimal performance on the test set.

Implementation For GRADE, we also utilize two GCN layers as the encoder. For hyperparameter settings, we vary the temperature τ in range $[0.5, 2]$, and the threshold ζ is searched in $[5, 15]$. The edge drop rate p_{edr} and feature drop rate p_{fdr} are tested in $[0.1, 0.4]$. We randomly initialize model parameters and use the Adam optimizer. Additionally, we employ random augmentation as a warmup since our graph augmentation relies on the quality of node representations. The number of epochs for a warmup is 200. The environment where we run experiments is:

- Operating system: Linux version 3.10.0-693.el7.x86_64
- CPU information: Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz
- GeForce RTX 3090

[5](MIT license) https://github.com/CRIPAC-DIG/GRACE
[6](MIT license) https://github.com/kavehhassani/mvgrl
[7](MIT license) https://github.com/hengruizhang98/CCA-SSG