ImGCL: Revisiting Graph Contrastive Learning on Imbalanced Node Classification

Liang Zeng[1], Lanqing Li[2*], Ziqi Gao[3], Peilin Zhao[2], Jian Li[1*]
[1]Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University
[2]Tencent AI Lab
[3]Hong Kong University of Science and Technology
zengl18@mails.tsinghua.edu.cn, {lanqingli, masonzhao}@tencent.com,
zgaoat@connect.ust.hk, lijian83@mail.tsinghua.edu.cn

Abstract

Graph contrastive learning (GCL) has attracted a surge of attention due to its superior performance for learning node/graph representations without labels. However, in practice, the underlying class distribution of unlabeled nodes for the given graph is usually imbalanced. This highly imbalanced class distribution inevitably deteriorates the quality of learned node representations in GCL. Indeed, we empirically find that most state-of-the-art GCL methods cannot obtain discriminative representations and exhibit poor performance on imbalanced node classification. Motivated by this observation, we propose a principled GCL framework on Imbalanced node classification (ImGCL), which automatically and adaptively balances the representations learned from GCL without labels. Specifically, we first introduce the online clustering based *progressively balanced sampling* (PBS) method with theoretical rationale, which balances the training sets based on pseudo-labels obtained from learned representations in GCL. We then develop the node centrality based PBS method to better preserve the intrinsic structure of graphs, by upweighting the important nodes of the given graph. Extensive experiments on multiple imbalanced graph datasets and imbalanced settings demonstrate the effectiveness of our proposed framework, which significantly improves the performance of the recent state-of-the-art GCL methods. Further experimental ablations and analyses show that the ImGCL framework consistently improves the representation quality of nodes in under-represented (tail) classes.

1 Introduction

Recently, graph contrastive learning (GCL) has become the *de facto standard* for self-supervised learning on graphs (Zhu et al. 2021a) due to its superior performance as compared to the supervised counterparts. (Thakoor et al. 2022; Bielak, Kajdanowicz, and Chawla 2021; You et al. 2020; Zhu et al. 2021b). Inheriting the advantage of self-supervised learning, GCL frees the model from the reliance on label information in the graph domain, where labels can be costly and error-prone in practice (Yang and Xu 2020) while unlabeled/partially labeled data is prevalent, such as fraudulent user detection (Kumar et al. 2018) and molecular property prediction (Ma et al. 2020). Typically, most GCL methods first construct multiple graph views via stochastic augmentation functions on the input graph and then learn discriminative representations by maximizing the representation consistency between two views.

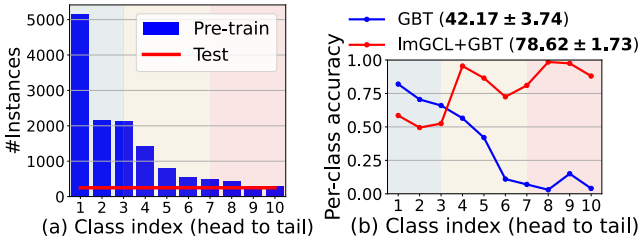


Figure 1: (a) Class distribution of the Amazon-Computers dataset sorted in decreasing order, where pre-training sets are highly *imbalanced* but testing sets are *balanced*. (b) Compared to the GBT (Bielak, Kajdanowicz, and Chawla 2021) baseline model on Amazon-computers, ImGCL+GBT substantially outperforms GBT on the overall accuracy. On three splits (*head/middle/tail*) depicted in different colors (*blue/yellow/red*), ImGCL+GBT improves the performance on middle and tail classes by a large margin with a slight sacrifice of accuracy on head classes.

Despite the prevalence and effectiveness of such methods, existing GCL methods mostly assume the datasets are carefully curated and well balanced across classes. However, unlabeled graph data randomly gathered in the wild often exhibits a highly imbalanced class distribution (Barabási and Albert 1999; Liu, Nguyen, and Fang 2021) and thus implicitly deteriorates the quality of learned representations in GCL methods (Jiang et al. 2021). For instance, Fig. 1(a) illustrates the class distribution sorted in decreasing order of the Amazon-Computers dataset (Shchur et al. 2018), a network of co-purchase goods containing 10 classes. In the pre-training set, we can clearly find that a small fraction of classes take up massive samples (*a.k.a.*, head classes) and the rest of classes are assigned only a few samples (*a.k.a.*, tail classes). Note that the highly *imbalanced* property of label information within the training data is unknown to the GCL methods. This important property implicitly exists and is largely ignored by the current GCL methods (Zhu et al. 2021a). However, to impose a fair eval-

*Corresponding authors.

uation metric, on imbalanced node classification, the testing set is *balanced* across all classes. Shown as the blue line of Fig. 1(b), we adopt one of the state-of-the-art GCL methods—GBT (Bielak, Kajdanowicz, and Chawla 2021)—to conduct experiments on imbalanced node classification. We can find that the baseline GBT model obtains very poor results, especially on the under-represented middle and tail classes, which naturally spurs a question:

How to improve the representation learning of GCL on highly imbalanced node classification?

Recent works related to this question (He et al. 2021; Kang et al. 2020a) explore balanced feature spaces to learn powerful representations not just for head classes but also for tailed classes. Kang et al. (2020b) introduce PBS method to innovate imbalanced representation learning, achieving remarkable success by decoupling the learning procedure into a representation learning stage and a classification stage. However, this method is impractical for the GCL setting since it requires knowing labels. This dilemma motivates us to explore how to implicitly obtain label information to improve node representations in the traditional GCL setting.

In this paper, we present a principled GCL framework on Imbalanced node classification (ImGCL), to automatically and adaptively balance the representations learned from GCL without ground truth labels. To perform class-balanced re-sampling, ImGCL obtains pseudo-labels via online clustering of learned representations in GCL. Moreover, we propose the node centrality based PBS method tailored for the graph domain, which assigns a higher probability to retain nodes with high node centrality scores when down-sampling the head class nodes in online clustering based PBS. This scheme is able to guide the model to learn node representations with higher node centrality scores, which is considered more important by having abundant structural connectivity when performing message passing. We also provide theoretical insight into the PBS method. Furthermore, existing GCL models can be seamlessly incorporated into the proposed ImGCL framework in a plug-and-play manner. In short, our main contributions can be summarized as follows:

- *New problem and insights:* we introduce a practically important but under-explored problem, namely graph contrastive learning on imbalanced node classification. We empirically identify that the recently proposed GCL methods are *vulnerable* to node class imbalance and result in large performance degradation.
- *New principled framework:* we propose a novel ImGCL framework, which utilizes the node centrality based PBS method. ImGCL automatically and adaptively balances the representations learned from GCL without knowing labels. Moreover, existing GCL models can be seamlessly incorporated into our framework.
- *Convincing empirical results:* we conduct comprehensive experiments to show that the ImGCL framework achieves superior performance compared with the recently proposed GCL methods on imbalanced node classification. Extensive ablation studies also demonstrate that the ImGCL framework improves the representations of the under-represented (middle and tail) classes.

2 Background

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{X})$ denote a graph, where $\mathcal{V} = \{v_i\}_{i=1}^N$ is a set of N nodes, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges between nodes. $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times d}$ represents the node feature matrix and $\boldsymbol{x}_i \in \mathbb{R}^d$ is the feature vector of node v_i, where d is the feature dimension. The adjacency matrix $\boldsymbol{A} \in \{0,1\}^{N \times N}$ is defined by $\boldsymbol{A}_{i,j} = 1$ if $(v_i, v_j) \in \mathcal{E}$ and 0 otherwise. More detailed discussions about related works can be found in the appendix.

Graph Contrastive Learning (GCL). Given an input graph, GCL aims to learn effective graph/node representations that can be transferred to downstream tasks by constructing positive and negative sample pairs (Thakoor et al. 2022; Bielak, Kajdanowicz, and Chawla 2021; Zhang et al. 2021a; Xu et al. 2021; Veličković et al. 2018; Sun et al. 2020; Hassani and Khasahmadi 2020). Specifically, we utilize two augmentation functions $t_1, t_2 \sim \mathcal{T}$ to generate graph views $\widetilde{\mathcal{G}}_1 = (\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{A}}_1) = t_1(\mathcal{G})$ and $\widetilde{\mathcal{G}}_2 = (\widetilde{\boldsymbol{X}}_2, \widetilde{\boldsymbol{A}}_2) = t_2(\mathcal{G})$, where $\mathcal{T}$ is the set of graph augmentation functions, such as node dropping, edge perturbation, and subgraph sampling (Zhu et al. 2021b). We then obtain node representations for the two graph views via the (parameter shared) GNN encoder $f(\cdot)$, denoted by $\boldsymbol{Z} = f(\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{A}}_1)$ and $\boldsymbol{Z}' = f(\widetilde{\boldsymbol{X}}_2, \widetilde{\boldsymbol{A}}_2)$ respectively. Given the latent representations, we optimize the parameters of the GNN encoder by a pre-defined contrastive loss. For any node u in $\widetilde{\mathcal{G}}_1$, we aim to score the positive pairs (u, u^+) higher compared to other negative pairs (u, u^-). Typically, the negative samples u^- are sampled from other nodes of the augmented graph views $\widetilde{\mathcal{G}}_2$ in the same batch. The commonly-used InfoNCE loss (Chen et al. 2020) can be defined as:

$$\mathcal{L}_{NCE}(u) = -\log \frac{s\left(\boldsymbol{z}_u, \boldsymbol{z}_{u^+}, \tau\right)}{s\left(\boldsymbol{z}_u, \boldsymbol{z}_{u^+}, \tau\right) + \sum_{u^- \neq u} s\left(\boldsymbol{z}_u, \boldsymbol{z}_{u^-}, \tau\right)}, \tag{1}$$

where $s\left(\boldsymbol{z}_u, \boldsymbol{z}_{u^+}, \tau\right)$ indicates the similarity between node representations of positive pairs, while $s\left(\boldsymbol{z}_u, \boldsymbol{z}_{u^-}, \tau\right)$ is the similarity between negative pairs. s is a contrasting function to measure the similarity between two node representations, which is typically defined as: $s(\boldsymbol{z}_u, \boldsymbol{z}_v, \tau) = \exp(\boldsymbol{z}_u \cdot \boldsymbol{z}_v / \tau)$. τ represents the temperature hyper-parameter.

Imbalanced Learning. Imbalanced learning seeks to learn a model from *the training set with an imbalanced class distribution*, where head classes take up the vast majority of samples and tail classes occupy only a few samples, and generalize well on *a balanced testing set* (Kang et al. 2020b; Zhang et al. 2021b,c; Liu et al. 2019; Li et al. 2022). For a K-way node classification problem, let $\{v_i, y_i\}_{i=1}^N$ be an imbalanced training set. The total number of the training set over K classes is $N = \sum_{k=1}^K N_k$, where N_k denotes the number of samples in class k. Let $\boldsymbol{\pi}$ be the vector of label frequencies, where $\pi_k = N_k/N$ denotes the label frequency of class k. Without loss of generality, we assume that the classes are sorted by π_k in a descending order (*i.e.*, if the class index $i < j$, then $N_i \geq N_j$, and $N_1 \gg N_K$). We denote by N_1/N_K the imbalance ratio of the dataset. Conven-

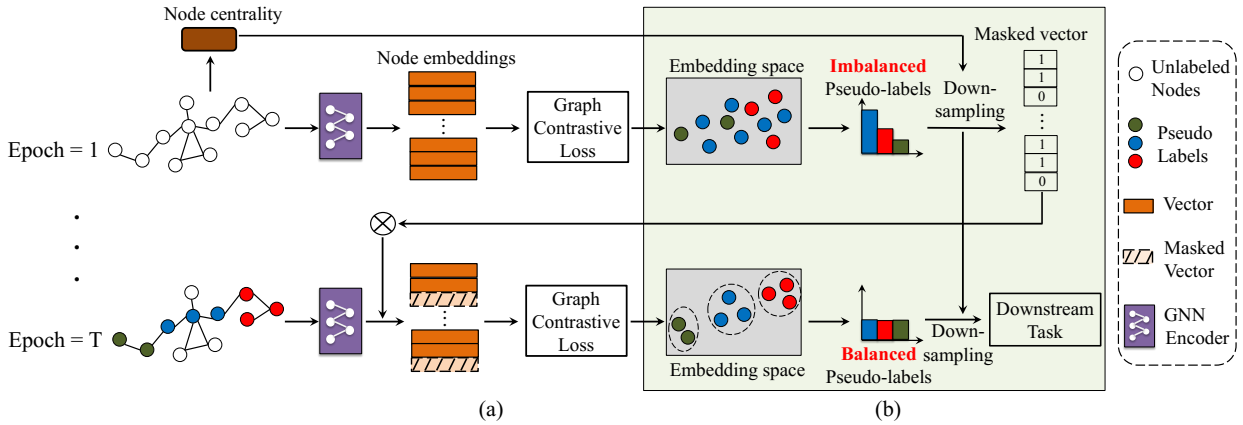


Figure 2: Overview of the proposed ImGCL framework. (a) Graph contrastive learning (GCL) methods take the graph as input and produce the embeddings of each node. (b) Node centrality based progressively balanced sampling (PBS) method automatically and adaptively balances the representations learned from GCL without knowing the labels.

tionally, evaluations on imbalanced learning report statistics for the head, middle, tail, and overall classes separately.

3 ImGCL: The Proposed Framework

In this section, we first introduce the progressively balanced sampling (PBS) method, which is impractical in our self-supervised setting since it requires knowing the labels. Motivated by the property of GCL, we generate pseudo-labels by clustering the node representations. Finally, we utilize the proposed node centrality based PBS method to adaptively attend to 'important' nodes of the given graph during the down-sampling phase. The overall framework of ImGCL is shown in Fig. 2.

3.1 Progressively Balanced Sampling (PBS)

Sampling Strategies. As suggested in (He et al. 2021), the probability p_k of sampling a node for the given graph from the class k is defined as:

$$p_k = \frac{N_k^q}{\sum_{i=1}^{K} N_i^q}, \tag{2}$$

where $q \in [0, 1]$. Different sampling strategies have different specific values of q.

PBS (Kang et al. 2020b). In imbalanced learning, the testing dataset is *balanced* whereas the training dataset is highly *imbalanced*. A single data sampling strategy which fits only one case, the balanced dataset ($p_k^M = \frac{1}{K}$ by setting $q = 0$ in Eq. 2) or the imbalanced dataset ($p_k^R = \frac{N_k}{\sum_{i=1}^{K} N_i}$ by setting $q = 1$ in Eq. 2), cannot account for the class distribution shift between the training and testing datasets. Thus, in order to learn high-quality representations from the imbalanced dataset, we adopt two data samplers with adaptive sampling strategies known as decoupled training in long-tailed learning literature (Zhang et al. 2021b). Formally, at training step t, data are sampled according to a linear combination of the random and mean strategies, controlled by a parameter $\alpha \in [0, 1]$. Therefore, the probability of sampling a node from the class k is given by:

$$\begin{aligned} p_k^{\text{PB}} &= \alpha * p_k^R + (1 - \alpha) * p_k^M \\ &= \alpha * \frac{N_k}{\sum_{i=1}^{K} N_i} + (1 - \alpha) * \frac{1}{K}. \end{aligned} \tag{3}$$

Intuitively, at early stages of the training phase, an imbalanced class distribution is used for representation learning of the feature extractor. At later stages, the model benefits more from a balanced dataset for training an unbiased classifier. Therefore, the control parameter α should progressively decrease from 1 to 0 during the training phase. Concretely, at training step t, α is calculated by (Kang et al. 2020b): $\alpha = 1 - \frac{t}{T}$, where T is the total number of training epochs.

3.2 Online Clustering Based PBS

The PBS method requires real labels to adjust the class distribution during training, which compromises its practicality. Since GCL is typically applied for the label-free scenario, we cannot directly apply PBS here. On one hand, (McPherson, Smith-Lovin, and Cook 2001) have revealed the homophily phenomenon in homophilic graphs, *i.e.*, the nodes with similar features tend to be connected with each other and share the same label. On the other hand, the motivation of GCL is to learn representations in which similar node pairs stay close to each other while dissimilar ones are far apart. We propose to connect these two facts by the pseudo-label method in GCL (Caron et al. 2018), which iteratively generates artificial labels by the model itself to make the PBS method applicable in the label-free GCL scenarios. We utilize the emergence of representation clusters learned from GCL to generate pseudo-labels of each node and then apply a down-sampling strategy to improve the quality of node representations in middle and tail classes.

Specifically, suppose there are K-classes for the node classification task. At the certain iteration t of the training phase, we obtain the node representation $\boldsymbol{Z}_t \in \mathbb{R}^{N \times D}$ via the learned GNN encoder, where D is the hidden dimension.

We apply the clustering algorithm to the nodes in the embedding space to produce a set of K prototypes $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$. Formally, we intend to learn a $D \times K$ centroid matrix C and a one-hot cluster assignment vector $\hat{y}_n \in \mathbb{R}_+^K$ for each node n of the given graph by solving the following problem:

$$\min_{C \in \mathbb{R}^{D \times K}} \frac{1}{N} \sum_{n=1}^{N} \min_{\hat{y}_n} \|\boldsymbol{z}_{t,n} - C\hat{y}_n\|_2^2 \text{ such that } \quad \hat{y}_n^\top 1_K = 1, \tag{4}$$

where $\boldsymbol{z}_{t,n} \in \mathbb{R}^D$ denotes the n-th node embedding vector of $\boldsymbol{Z}_t$, and $1_K \in \mathbb{R}^K$ is the vector with all elements of 1. Solving this above problem provides a set of optimal assignments $\{\hat{y}_n^* | n = 1, \ldots, N\}$ and a centroid matrix C^*. These assignments are then used as pseudo-labels. Note that *We set the number of centroids (clusters) K in the clustering algorithm equal to the number of classes in the training dataset*, which is an input hyperparameter of the classification task. We also perform the hyperparameter study in Appendix D and find that it is generally a good choice and prevents performance fluctuation. In order to avoid trivial solutions and empty clusters, we use the *constrained K-means clustering* (Bradley, Bennett, and Demiriz 2000) to instantiate the clustering algorithm. It can implement the K-means clustering algorithm whereby a minimum size for each cluster can be specified. Thus, we can address the representation collapse problem (Fang et al. 2021) which would produce a highly imbalanced pseudo-label distribution.

Theoretical Analysis. In order to justify the PBS method, we theoretically prove that, the classifier learned iteratively by balanced sampling with pseudo-labels on the imbalanced dataset can converge to the optimal balanced classifier with a linear rate. Detailed proofs can be found in Appendix B.

Consider a binary classification problem with two Gaussian distributions with different means and the equal variance. Suppose the data generating distribution is P_{XY} and the probability of positive labels $(+1)$ and negative labels (-1) are $P_Y(1)$ and $P_Y(-1)$, respectively. We have $X|Y = +1 \sim \mathcal{N}(\mu_1, \sigma^2)$ conditioned on $Y = +1$ and similarly, $X|Y = -1 \sim \mathcal{N}(\mu_2, \sigma^2)$ conditioned on $Y = -1$. In addition, suppose $\mu_1 < \mu_2$ without loss of generality. It is straightforward to verify that (Bishop and Nasrabadi 2006) the optimal decision boundary of a balanced Bayes classifier is $\theta^* \equiv \frac{\mu_1+\mu_2}{2}$. At the first iteration $t = 1$, we start from the imbalanced unlabeled dataset and generate pseudo-labels $\hat{Y}_0$ by the clustering method. Then, we obtain the estimated decision boundary $\hat{\theta}_1 = \frac{\hat{\mu}_1+\hat{\mu}_2}{2}$ and produce pseudo-labels $\hat{Y}_1$. The fact that the initial dataset being imbalanced $(P_Y(1) \neq P_Y(-1))$ leads to biased $\hat{\theta}_1$. We iteratively obtain the estimator $\hat{\theta}_t$ and pseudo-labels $\hat{Y}_t$ when $t \geq 2$.

Proposition 1. *Consider the above setup. Suppose there is a (sufficiently large) integer T such that $|\hat{\theta}_T - \theta^*| \ll |\mu_2 - \mu_1|$, $|(\hat{\theta}_T - \theta^*)(\mu_2 - \mu_1)| \ll \sigma^2$. Our estimator $\hat{\theta}_T$ converges to the optimal balanced decision boundary θ^*, i.e., $|\hat{\theta}_{t+1} - \theta^*| \leq C \cdot |\hat{\theta}_t - \theta^*|, \forall t \geq T$ with a linear convergence rate $C = \frac{2}{\pi} < 1$.*

Interpretation. In the context of PBS, the classifier is trained starting from the imbalanced data distribution. Intuitively, by iteratively down-sampling head classes, the data distribution gradually becomes balanced, on which the trained classifier also converges to the balanced optimum.

3.3 Node Centrality Based PBS

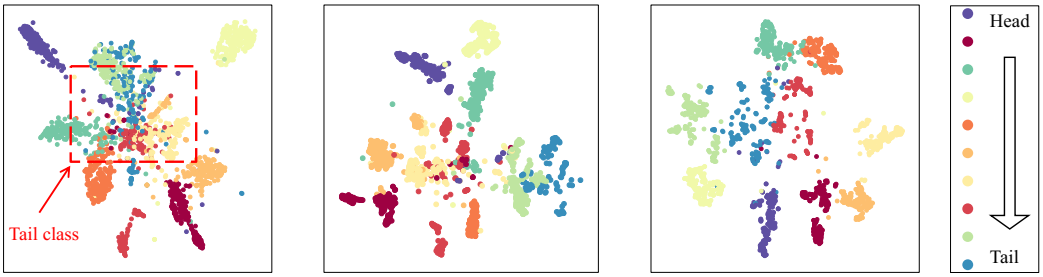
To further improve the representations of nodes in under-represented (middle and tail) classes, we incorporate graph structural information when performing the node centrality based PBS method. In network science, node centrality, which measures how important a node is in a graph, is an important metric to understand the influence of each node of a graph (Barabási 2013). Therefore, we propose an adaptive down-sampling scheme based on node centrality to balance the class distribution over all classes. For each node class, we sample nodes of higher centrality with higher probability to better preserve the intrinsic structures of graphs in learned representations. We herein utilize the PageRank centrality due to its simplicity and effectiveness (Barabási 2013). Formally, the centrality values are calculated by the iterative form: $\boldsymbol{\sigma} = \alpha \boldsymbol{A}\boldsymbol{D}^{-1} + \mathbf{1}$, where $\boldsymbol{\sigma} \in \mathbb{R}^N$ is the PageRank centrality score vector for each node, α is a damping factor to control the probability of randomly jumping to another node in the graph, A and D denote the adjacency and the degree matrix of the input graph respectively, and $\mathbf{1}$ is the all-ones identity vector. Note that we pre-calculate the PageRank score σ before the training phase of GCL. When performing down-sampling of the head classes, we calculate the probability of each node based on σ. Formally, for node v in class j with the centrality score σ_v, the probability with the normalized centrality score is defined as:

$$p_{v,j}^{\text{NPB}} = \max\left\{ \frac{\boldsymbol{\sigma}_v - \boldsymbol{\sigma}_{\min}}{\boldsymbol{\sigma}_{\max} - \boldsymbol{\sigma}_{\min}} \cdot p_j^{\text{PB}}, p_\tau \right\}, \tag{5}$$

where p_j^{PB} is the progressively balanced sampling probability of class j, $\boldsymbol{\sigma}_{\max}$ and $\boldsymbol{\sigma}_{\min}$ are the maximum and minimum value of the centrality score, and p_τ is a cut-off probability to ensure that nodes with extremely low probabilities can also be sampled. In node centrality based PBS, we then perform a normalization step that transforms $p_{v,j}^{\text{NPB}}$ into probabilities and then use it to balance the class distribution. Concretely, we select certain nodes of the original graph in form of a *masked vector* $\boldsymbol{m} \in \mathbb{R}^N$ by sampling each node independently according to $p_{v,j}^{\text{NPB}}$. We then calculate the graph contrastive loss only on these selected node representations, as shown in Fig. 2.

3.4 Learning Framework

ImGCL is a general GCL framework for imbalanced node classification, which can be readily applied with existing GCL methods adopting the two-branch design (Thakoor et al. 2022; Bielak, Kajdanowicz, and Chawla 2021; You et al. 2020). ImGCL does not rely on specific approaches of graph view augmentation, graph view encoding, or representation contrasting in GCL. In ImGCL, we set the number of clusters K in the node centrality based PBS method equal to the number of classes in the downstream task. One



(a) **GBT** (b) **GBT+PBS** (c) **ImGCL+GBT**

Figure 3: Visualization of the testing set on Amazon-Computers. Each point in the figure is colored by real labels.

challenge of the implementation is that the quality of the generated pseudo-labels and representations are mutually-dependent, which could destabilize the training loop. In response, we re-balance the class distribution every B epochs, thus there are T/B times to adjust the class distribution. In order to improve the representations of nodes in under-represented (middle and tail) classes, we select $N \times l$ nodes during the pre-training phase in ImGCL, where $l = 10\%$ equals the ratio of training data in the down-stream task. We down-sample the head nodes to the required number according to their PageRank centrality scores as introduced in Sec. 3.3. Following the linear evaluation scheme of GCL (Zhu et al. 2021b), we train a linear classifier on the balanced dataset with 10% randomly selected nodes after obtaining node representations. The training algorithm of ImGCL is summarized as follows.

Algorithm 1: The ImGCL pre-training algorithm

Input: The input graph $\mathcal{G}$, GNN encoder $\mathcal{F}$.
Parameter: Number of nodes N, number of clusters K, re-balanced labeling frequency B, the ratio of selected nodes l.
Output: Pre-trained GNN encoder $\mathcal{F}$.

Calculate the node centrality vector σ of $\mathcal{G}$.
for $epoch = 0, 1, 2, \cdots$ **do**
  Draw two augmentation functions $t_1 \sim \mathcal{T}$, $t_2 \sim \mathcal{T}$.
  Generate two graph views $\widetilde{\mathcal{G}}_1 = t_1(\mathcal{G})$ and $\widetilde{\mathcal{G}}_2 = t_2(\mathcal{G})$.
  Obtain node representations U of $\widetilde{\mathcal{G}}_1$ and V of $\widetilde{\mathcal{G}}_2$ using the GNN encoder $\mathcal{F}$.
  if $epoch \ mod \ B == 0$ **then**
    Cluster node representations to obtain pseudo-labels.
    Calculate the normalized centrality score p^{NPB} with Eq. 5.
    Obtain the masked vector m according to p^{NPB}, which satisfies $\|m\| = N \times l$.
  Compute the contrastive object $\mathcal{L}$ on these selected node representations $U \odot m$ and $V \odot m$ with Eq. 1.
  Update the parameters of $\mathcal{F}$ with $\mathcal{L}$.

4 Case Study: Learning Discriminative Representations on Amazon-Computers

We are particularly interested in the learned representations among different methods. For a more intuitive comparison and to further demonstrate the effectiveness of the ImGCL framework, we design experiments of visualization on Amazon-computers. We utilize the output representations on the last layer of vanilla GBT, GBT with PBS to train the classifier on training data but without node centrality based PBS on representation learning in GCL, and GBT with ImGCL. We plot the learned representations of the testing graph dataset using t-SNE (Van der Maaten and Hinton 2008). As shown in Fig. 3, vanilla GBT clearly exhibits the minority collapse (Fang et al. 2021) phenomenon, and the representations in tail classes are mixed together, which fundamentally limits the performance in the tail classes. In Fig. 3 (b), we can find clearer boundaries among different classes but the representations in tail classes are still mixed together, which suggests that it is important to explore balanced representation spaces in GCL methods. In comparison, the learned representations of GBT+ImGCL (Fig. 3 (c)) have distinct boundaries among different classes and more compact intra-class structures, which highlights the effectiveness of our proposed ImGCL framework.

5 Experiments

In this section, we provide empirical results to demonstrate the effectiveness of our ImGCL framework. We conduct extensive experiments on imbalanced graph datasets to mainly answer the following questions: [1] (1) Can ImGCL generally improve the performance of GCL methods on imbalanced node classification? (Sec. 5.1) (2) How does ImGCL perform on different imbalanced types? (Sec. 5.2) (3) How does ImGCL help improve GCL methods on imbalanced node classification? (Sec. 5.2)

Dataset. We use four widely-used datasets including Wiki-CS, Amazon-computers, Amazon-photo, and DBLP, to comprehensively study the performance of transductive node classification. In order to validate the effectiveness of ImGCL on the imbalanced node classification setting, we

[1] Due to space limitations, ablations on different components in ImGCL and hyperparameter studies are provided in the appendix.

Category	Method	Available Data	Amazon-Computers	Amazon-Photo	Wiki-CS	DBLP
	Raw features	$\boldsymbol{X}$	33.99(4.47)	38.07(4.03)	34.50(2.06)	38.51(0.80)
	Node2vec	$\boldsymbol{A}$	69.80(2.69)	69.44(0.38)	**51.76(2.24)**	50.41(2.77)
	DeepWalk	$\boldsymbol{A}$	69.67(2.36)	69.00(0.62)	51.32(2.17)	**50.57(2.88)**
	DeepWalk + features	$\boldsymbol{X}, \boldsymbol{A}$	**70.20(3.30)**	**71.60(3.31)**	51.51(2.51)	49.57(1.65)
GCL	DGI	$\boldsymbol{X}, \boldsymbol{A}$	10.88(2.09)	13.62(2.26)	17.11(4.83)	26.63(10.87)
	MVGRL	$\boldsymbol{X}, \boldsymbol{A}$	13.40(3.01)	16.92(3.14)	45.97(2.42)	44.43(0.57)
	InfoGraph	$\boldsymbol{X}, \boldsymbol{A}$	35.83(7.01)	53.57(13.29)	44.19(3.90)	48.26(5.95)
	GRACE	$\boldsymbol{X}, \boldsymbol{A}$	**41.54(2.51)**	45.24(4.24)	**54.20(3.97)**	44.48(0.40)
	BGRL	$\boldsymbol{X}, \boldsymbol{A}$	40.81(5.01)	51.18(10.49)	39.82(3.60)	49.58(3.99)
	GBT	$\boldsymbol{X}, \boldsymbol{A}$	42.17(3.74)	**60.73(3.64)**	44.95(2.63)	**58.51(5.30)**
ImGCL (ours)	DGI	$\boldsymbol{X}, \boldsymbol{A}$	48.85(10.94)	47.99(9.03)	41.20(18.84)	50.39(9.17)
	MVGRL	$\boldsymbol{X}, \boldsymbol{A}$	46.42(11.33)	50.86(9.49)	60.85(2.60)	51.90(7.42)
	InfoGraph	$\boldsymbol{X}, \boldsymbol{A}$	75.44(5.30)	72.56(2.91)	68.96(5.86)	69.12(3.24)
	GRACE	$\boldsymbol{X}, \boldsymbol{A}$	77.54(3.00)	68.89(4.41)	**73.86(2.78)**	63.61(4.91)
	BGRL	$\boldsymbol{X}, \boldsymbol{A}$	67.82(10.24)	72.67(6.04)	59.35(15.44)	57.90(2.28)
	GBT	$\boldsymbol{X}, \boldsymbol{A}$	**78.62(1.73)**	**75.13(7.13)**	73.08(12.45)	**70.05(1.78)**
Best ImGCL over GCL			**39.61** ↑	**34.47** ↑	**28.13** ↑	**23.76** ↑
Supervised	GCN	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	46.83(1.52)	68.84(2.56)	59.84(2.02)	51.55(2.64)
	GCN+PBS	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	**70.12(9.78)**	**73.34(8.28)**	**63.15(5.13)**	**73.11(2.76)**

Table 1: Summary of accuracies (%) with standard deviation on imbalanced node classification. The 'Available Data' means data we can obtain for training, where $\boldsymbol{X}$, $\boldsymbol{A}$, and $\boldsymbol{Y}$ denote node features, the adjacency matrix, and labels respectively. We highlight models in the ImGCL category with the gray background. The highest performance under each category is masked as bold. The highest performance improvement of the GCL baseline w & w/o the ImGCL framework is underlined.

select an equal number of nodes in each class for the validation and testing dataset. Following (Zhu et al. 2021b), the training set is randomly sampled from the rest according to train/valid/test ratios = 1:1:8, which is highly imbalanced. The descriptions, statistics, and the imbalance ratio of each dataset can be found in Appendix H.

Evaluation Protocol. For each experiment, we follow the commonly-used *linear evaluation scheme* for GCL as introduced in (Zhu et al. 2021b). The model is firstly trained in a self-supervised manner, and then the learned representations are used to train and test with a simple linear classifier. For results in this section, we train each model in twenty runs for different data splits and report the average performance with the corresponding standard deviation for a fair comparison. In what follows, we measure performance in terms of accuracy, if not otherwise specified.

Imbalanced Types. In order to comprehensively evaluate the performance of ImGCL in different imbalanced types, we introduce two imbalanced types (Jiang et al. 2021): *Exp* and *Pareto*, parameterized by an imbalanced factor. *Exp* imbalanced class distribution is given by an exponential function, where the higher imbalanced factor means the more imbalanced graph. *Pareto* imbalanced class distribution is determined by a Pareto distribution, where a lower imbalanced factor means the smaller power value of a Pareto distribution and thus the more imbalanced graph.

Baselines. We consider representative baseline methods in the following two categories: (1) traditional methods including Node2vec (Grover and Leskovec 2016), DeepWalk (Perozzi, Al-Rfou, and Skiena 2014), and raw features as input without considering the graph topology. (2) deep learning methods including DGI (Veličković et al. 2018), MVGRL (Hassani and Khasahmadi 2020), InfoGraph (Sun et al. 2020), GRACE (Zhu et al. 2021b), BGRL (Thakoor et al. 2022), and GBT (Bielak, Kajdanowicz, and Chawla 2021). We also directly compare ImGCL with the supervised counterparts, *i.e.*, the most representative model GCN (Kipf and Welling 2016) and the variant of GCN trained with PBS. Note that for all baselines, we report their performance on the imbalanced experimental settings following their official hyperparameters (detailed in Appendix H) based on the PyGCL (Zhu et al. 2021a) open-source library.

5.1 Experimental Results on Node Classification

The empirical performance of imbalanced node classification with the *Exp* type and 100 imbalanced factor is summarized in Table 1. ImGCL consistently outperforms current GCL baselines or even the supervised baselines. We summarize our observations from the table as follows: (1) Recently proposed GCL methods (Zhu et al. 2021a), which are evaluated on balanced testing sets, exhibit severe performance degradation in our imbalanced node classification setting. By incorporating our proposed ImGCL framework (the gray background), these GCL methods improve by a large margin. Concretely, ImGCL+GCL achieves [39.61%, 34.47%, 28.13%, 23.76%] average absolute gain in accuracy than the baseline GCL models on [Amazon-Computers, Amazon-Photo, Wiki-CS, DBLP], respectively. We also find that the recently proposed GBT (Bielak, Kajdanowicz, and Chawla 2021) obtains the best performance among a set of GCL competitors. We think the reason is that *feature decorrela-*

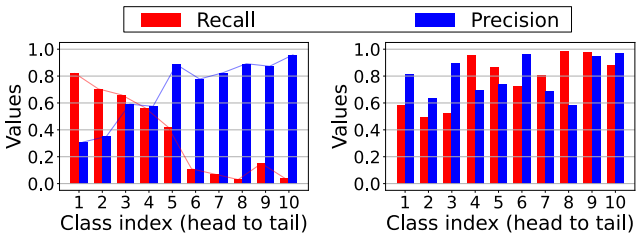
Type	Factor	Method Catergory	Head	Middle	Tail	All
Exp ↑	20	GBT	79.45	65.40	73.25	71.97
	50		80.12	60.57	57.42	65.49
	100		70.38	41.92	14.29	42.17
	200		78.13	25.12	10.28	36.57
	20	GBT+ImGCL (ours)	71.58	89.95	94.55	85.82
	50		63.88	86.03	95.75	82.30
	100		57.32	82.72	94.45	78.62
	200		46.95	75.27	76.42	67.12
Pareto ↓	2	GBT	83.31	61.01	52.58	65.17
	1		82.06	57.22	39.85	59.46
	2	GBT+ImGCL (ours)	53.81	70.25	93.19	72.20
	1		41.94	78.34	87.97	70.31

Table 2: Results of accuracy (%) on Amazon-computers using the GBT baseline model under different imbalanced types and factors. ↑ means a higher imbalanced factor corresponds to a more imbalanced dataset. Instead, ↓ means a lower factor corresponds to a more imbalanced dataset.

tion method in GBT is more fit for the imbalanced node classification and we adopt the GBT baseline model in the following experimental analysis. (2) The traditional methods, *e.g.*, Node2vec and DeepWalk, can achieve competitive performance in the imbalanced node classification task compared with GCL methods. We postulate the reason is that the traditional network embedding methods can take advantage of the homophily (Barabási 2013) property to utilize the graph topology features which is important on imbalanced node classification, as reflected in (Liu, Nguyen, and Fang 2021). Moreover, the "Raw features" method without considering the graph topology cannot achieve satisfactory performance, which indicates the necessity of utilizing the graph topology features on imbalanced node classification. (3) Compared with the supervised learning methods GCN and GCN+PBS, the ImGCL framework achieves superior or competitive performance on all datasets, which further corroborates the effectiveness of our proposed framework.

5.2 Experimental Analysis

Imbalanced Type Analysis. The performance for the *head, middle, and tail* classes are usually reported on imbalanced learning (Zhang et al. 2021b). We first divide the training set of Amazon-computers into three disjoint groups in terms of class size: $\{$*Head*, *Middle*, *Tail*$\}$. *Head* and *Tail* each include the top and bottom $1/3$ classes, respectively. Because there are 10 classes of Amazon-computers in total, the classes with sorted indices in decreasing order [1-3, 4-7, 8-10] belong to [*Head* (3 classes), *Middle* (4 classes), *Tail* (3 classes)] categories, respectively. To validate ImGCL across different imbalanced class distributions, we design experiments on Amazon-computers using the GBT baseline model w & w/o the ImGCL framework. We consider two imbalanced types: Exp and Pareto. Four imbalanced factors [20, 50, 100, 200] are associated with Exp. Two imbalance factors [1, 2] are chosen for Pareto. In Table 2, we observe that the more imbalanced dataset leads to the lower accuracy



(a) **GBT** (b) **ImGCL+GBT**

Figure 4: Bias comparison between GBT models w & w/o ImGCL on Amazon-Computers under imbalanced experimental settings. Left: Per-class recall and precision w/o ImGCL. Right: Per-class recall and precision with ImGCL. The class index is sorted by the number of nodes in each class in descending order. GBT w/o ImGCL clearly shows a descending trend in recall while an ascending trend in precision. However, GBT with ImGCL has achieved a relatively balanced recall and precision value over all classes.

of the model. Nevertheless, according to our assumption in Proposition 1, the ImGCL+baseline model consistently outperforms the baseline model.

Per-Class Analysis. We analyze the per-class recall and precision in Fig. 4 to better understand how ImGCL can help improve GCL methods on the imbalanced node classification. We test GBT on Amazon-computers under imbalanced experimental settings introduced in Sec. 5. The GBT baseline model exhibits highly skewed performance on head and tail classes. The recall on the most majority and minority class is 82.0% and 4.1% respectively, while the corresponding numbers for precision are 30.9% and 95.8%. We observe that GBT falsely classifies most of the tail class samples into head classes with low confidence but have high confidence on the correctly classified nodes in tail classes (Wei et al. 2021). In contrast, GBT with ImGCL obtains relatively balanced recall and precision value on both head (the most majority precision: 30.9% $\rightarrow$ 81.2%) and tail (the most minority recall: 4.1% $\rightarrow$ 88.2%) classes, which leads to the substantial improvement in the overall accuracy (*i.e.*, 42.17% $\rightarrow$ 78.62%) across all classes, as shown in Table 3.

6 Conclusion

In this work, we study how to improve the representations of graph contrastive learning (GCL) methods on imbalanced node classification, which is a very practical but rarely explored problem. We propose the principled ImGCL framework, which automatically and adaptively balances the representations learned from GCL without knowing labels and then theoretically justifies it. Through extensive experiments on multiple graph datasets and imbalance settings, we show that ImGCL can significantly improve the recently proposed GCL methods by improving the representations of nodes in under-represented (tail) classes. For the future work, we will explore more data types, such as bioinformatics graphs. We hope our work will extend GCL to more realistic task settings with (underlying) imbalanced node class distribution.

Acknowledgments

Liang Zeng and Jian Li are supported in part by the National Natural Science Foundation of China Grant 62161146004, Turing AI Institute of Nanjing and Xi'an Institute for Interdisciplinary Information Core Technology.

References

Barabási, A.-L. 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987): 20120375.

Barabási, A.-L.; and Albert, R. 1999. Emergence of scaling in random networks. *science*, 286(5439): 509–512.

Bielak, P.; Kajdanowicz, T.; and Chawla, N. V. 2021. Graph Barlow Twins: A self-supervised representation learning framework for graphs. *arXiv preprint arXiv:2106.02466*.

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Bradley, P. S.; Bennett, K. P.; and Demiriz, A. 2000. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0): 0.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 132–149.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, 1597–1607. PMLR.

Fang, C.; He, H.; Long, Q.; and Su, W. J. 2021. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43).

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 855–864.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning (ICML)*, 4116–4126. PMLR.

He, J.; Kortylewski, A.; Yang, S.; Liu, S.; Yang, C.; Wang, C.; and Yuille, A. 2021. Rethinking Re-Sampling in Imbalanced Semi-Supervised Learning. *arXiv preprint arXiv:2106.00209*.

Jiang, Z.; Chen, T.; Mortazavi, B.; and Wang, Z. 2021. Self-Damaging Contrastive Learning. In *International Conference on Machine Learning (ICML)*.

Kang, B.; Li, Y.; Xie, S.; Yuan, Z.; and Feng, J. 2020a. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations (ICLR)*.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020b. Decoupling representation and classifier for long-tailed recognition. In *8th International Conference on Learning Representations (ICLR)*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kumar, S.; Hooi, B.; Makhija, D.; Kumar, M.; Faloutsos, C.; and Subrahmanian, V. 2018. Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*, 333–341.

Li, L.; Zeng, L.; Gao, Z.; Yuan, S.; Bian, Y.; Wu, B.; Zhang, H.; Lu, C.; Yu, Y.; Liu, W.; et al. 2022. ImDrug: A Benchmark for Deep Imbalanced Learning in AI-aided Drug Discovery. *arXiv preprint arXiv:2209.07921*.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2537–2546.

Liu, Z.; Nguyen, T.-K.; and Fang, Y. 2021. Tail-GNN: Tail-Node Graph Neural Networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, 1109–1119.

Ma, H.; Bian, Y.; Rong, Y.; Huang, W.; Xu, T.; Xie, W.; Ye, G.; and Huang, J. 2020. Multi-View Graph Neural Networks for Molecular Property Prediction. *arXiv preprint arXiv:2005.13607*.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1): 415–444.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 701–710.

Shchur, O.; Mumme, M.; Bojchevski, A.; and Günnemann, S. 2018. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*.

Sun, F.-Y.; Hoffmann, J.; Verma, V.; and Tang, J. 2020. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Thakoor, S.; Tallec, C.; Azar, M. G.; Munos, R.; Veličković, P.; and Valko, M. 2022. Large-Scale Representation Learning on Graphs via Bootstrapping. In *International Conference on Learning Representations (ICLR)*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Veličković, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2018. Deep graph infomax. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wei, C.; Sohn, K.; Mellina, C.; Yuille, A.; and Yang, F. 2021. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10857–10866.

Xu, D.; Cheng, W.; Luo, D.; Chen, H.; and Zhang, X. 2021. InfoGCL: Information-Aware Graph Contrastive Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34.

Yang, Y.; and Xu, Z. 2020. Rethinking the value of labels for improving class-imbalanced learning. In *Thirty-Fourth Advances in Neural Information Processing Systems (NeurIPS)*.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. In *Advances in neural information processing systems (NeurIPS)*.

Zhang, H.; Wu, Q.; Yan, J.; Wipf, D.; and Yu, P. S. 2021a. From canonical correlation analysis to self-supervised graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2021b. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*.

Zhang, Y.; Wei, X.-S.; Zhou, B.; and Wu, J. 2021c. Bag of Tricks for Long-Tailed Visual Recognition with Deep Convolutional Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3447–3455.

Zhu, Y.; Xu, Y.; Liu, Q.; and Wu, S. 2021a. An Empirical Study of Graph Contrastive Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2021b. Graph Contrastive Learning with Adaptive Augmentation. In *Proceedings of the Web Conference (WWW)*, 2069–2080.