Gradient Recalibration for Improved Visibility of Tail Classes in Supervised Contrastive Learning

1st Genze Zhan
Beijing institute of technology
Beijing, China
genzezhan@bit.edu.cn

2th Xin Li
Beijing Institute of Technology
Beijing, China
xinli@bit.edu.cn

3nd Yong Heng
Beijing Institute of Electronic System Engineering
Beijing, China
hengyong_BIESE@163.com

4rd Yan Zhang*
Beijing Institute of Technology
Beijing, China
7420180300@bit.edu.cn

5th Jiaojiao Wang
Institute of Automation Chinese Academy of Sciences
Beijing, China
jiaojiao.wang@ia.ac.cn

6th Peiyao Zhao
Beijing Institute of Technology
Beijing, China
peiyaozhao@bit.edu.cn

7th Meitao Mu
Beijing Institute of Technology
Beijing, China
3120220960@bit.edu.cn

8th Xueying Zhu
Beijing Institute of Technology
Beijing, China
3120230893@bit.edu.cn

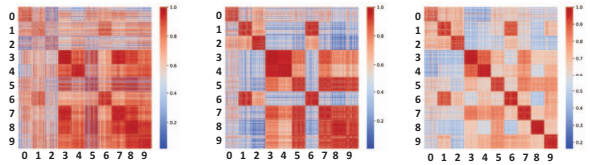
9th Mingzhong Wang
The University of the Sunshine Coast
Queensland, Australia
mwang@usc.edu.au

***Abstract*—Contrastive learning and supervised contrastive learning (SCL) have proven their effectiveness in graphs. However, they suffer from representation collapse when meet imbalance. To address these, we first proposed a quantitative model, similar to the Thomson problem when all classes are of equal size. It maps classes on the hypersphere where different classes repel each other. Based on this, we theoretically showed that when applied to imbalanced node classification, tail classes will be pushed together due to the dominating repellent forces from head classes. Therefore, we recalibrate the gradient of SCL loss to enforce all classes to maintain a uniform distribution in feature space, improving the visibility of tail classes. Extensive experiments on graph datasets indicates that the proposed method can significantly enhance the uniformity of class representation, thus achieving better performance for imbalanced node classification.**

***Index Terms*—graph neural network, supervised contrastive learning, Thomson problem, representation collapse**

I. Introduction

Graph contrastive learning (GCL), for graph-structured data, is a representations learning technique by pulling positive pairs close and pushing negative pairs away in the embedding space. Existing GCL algorithms generally break the underlying semantic structures due to the uniformity of the embedding distribution [1]. SCL [1] is then proposed to better preserve semantic information by leveraging label information. However, in real-life scenarios, graph datasets are usually imbalanced [2]. When meet this situation, they suffer from representation collapse and tend to map nodes from tail classes into similar representations [3]. Fig. 1 depicts the heat maps of node-wise similarity based on the representations learned by three models on an imbalanced dataset, including GRACE [4], GRACE (SCL)which replaces the CL loss with SCL loss, and our proposed approach (OURS). The X-axis represents the class index, reflecting the size of each class in descending order. Fig. 1(a) shows that the similarity within a class is not well revealed, which is then alleviated by leveraging the SCL loss instead, as shown in Fig. 1(b). However, the high similarity between the tail classes(the three rightmost classes on the X-axis), indicates representation collapse, which maps different data into the same representation, making it difficult for a classifier to separate samples into their respective classes.



(a) GRACE (b) GRACE (SCL) (c) OURS

Fig. 1. The heat maps of node similarity matrices in the feature space of GRACE, GRACE (SCL), and OURS on the Amazon-Computers dataset.

To solve this problem, we propose the Gradient-Recalibrated SCL Loss (GR-SCL Loss) to avoid representation entanglement for tail classes. We first model classes as repellent electrons on a unit hyper-sphere to quantitatively analyze the impact of class sizes on the representation distribution learned from SCL. Secondly we start with the solutions provided by the Thomson Problem [5] as the stable state of representation in the balance setting and deduct the representation collapse in the imbalance setting. Then, we reduce the gradient from head classes and increase the gradient

*Corresponding author

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00126

from tail classes. Finally, extensive experiments demonstrate that with the Loss the distribution of features from different classes becomes more uniform, as Fig. 1(c) shows.

Contributions:(1)We proposed a quantitative model to evaluate the impact of class sizes on representation distribution, which resembles the Thomson problem. Based on that, we theoretically showed that when applied to imbalanced node classification, SCL suffers from representation collapse. (2)We designed a novel GR-SCL Loss for imbalance setting, which can learn the discriminative embedding to avoid representation collapse. (3)We experimented on benchmark datasets to show the validity and superiority of our method compared to the state-of-the-art imbalanced node classification competitors.

II. Related Work

The majority of exsting methods for graph imbalance learning can be basically regarded as sample-based, and mainly consider at the data level. Reference [6] attempted to oversample tail classes to balance the datasets. However, this type of methods usually overfits tail and hurts the performance of the head [7]. GraphSMOTE [8] synthesizes nodes in the feature space for tail classes, but the process of calculating the nearest node pairs is time-consuming. DR-GCN [9] introduces a class-conditioned adversarial training process and pursues unlabeled and labeled node consistency. ImGAGN [10] train a generator for the tail to balance the distribution. Its extension to multi-class classification, unfortunately, disregards sub-tail classes, resulting in suboptimal performance. DPGNN [11] uses a class prototype-driven balance training scheme to transfer knowledge from head to tail. On datasets with many classes, however, the dimension of the concatenated embedding difference becomes excessively high, resulting in poor performance due to the curse of high dimensionality [12]. This paper is considered from representation learning and explores the application of graph contrast learning in imbalanced datasets.

III. Preliminary

A. Problem Formulation

Let $G = (V, \mathcal{E}, X)$ denote an attributed graph, where V is the set of n nodes, $\mathcal{E}$ is the set of edges and $X \in R^{n\times f}$ is the node attribute matrix. $\mathcal{C} = \{0, 1, ..., C-1\}$ is the set of C classes, and $\mathcal{Y} \in R^{n\times C}$ is the one-hot node label matrix, where y_i denotes the label of v_i. For convenience, we uniformly use lowercase letters with arrows to denote node representations. Let $\{N_0, ..., N_{C-1}\}$ be the number of nodes for corresponding classes. The classes are sorted by their cardinality in decreasing order (i.e., if $i_1 < i_2$, then $N_{i_1} \geq N_{i_2}$) [7, 13]. The imbalance ratio ρ is defined as $\frac{N_0}{N_{C-1}}$ [14, 15], indicating the degree of imbalance in the dataset. With the notations above, the imbalanced node classification is formulated as: *Given a graph G with imbalanced node classes($\rho >> 1$), the goal is to learn a boundary-clear feature extractor g which outputs $H \in R^{n\times d}$ as the node embedding matrix, which remarkably reduces classification errors against tail classes. Formally,*

$$g(G[, \mathcal{Y}]) = H.$$

With $\mathcal{Y}$ considered, it becomes a SCL problem.

B. Conventional SCL Loss

Contrastive learning (CL) generally uses augmented views as the only positive samples and ignores other potential positive samples. Therefore, semantically similar samples from the same class are pulled apart, which hurts the representation alignment. SCL takes instances from the same class into consideration as positive samples. Given an labeled embedding set $\{\overrightarrow{h_i}\}_{i=1}^{n}$, the SCL loss of node v_i is formulated as

$$L_i = -\frac{1}{|P(i)|}\sum_{p\in P(i)} log\frac{exp(S_{ip}/\tau)}{\sum_{k\in V(i)\cup P(i)} exp(S_{ik}/\tau)}, \tag{1}$$

where $S_{ij} = \overrightarrow{h_i}^T\overrightarrow{h_j}$ is the similarity between the embeddings of the node v_i and node v_j. $P(i) = \{v_p \in V : y_p = y_i\}$ is the set of positive samples with the same label as v_i, while $V(i) = \{v_k \in V : y_k \neq y_i\}$ is the set of negative samples with different labels from v_i. τ is a temperature hyper-parameter. $\overrightarrow{h_i}$ is the normalized representation: $\forall i \in V, ||\overrightarrow{h_i}|| = 1$.

IV. Gradient Analysis for Balanced Data

This section analyzes the gradient-based representation updates for balanced data, focusing on investigating how negative samples contribute to representation updates for convenience.

For node v_i, the gradient of its SCL loss L_i (1) with respect to the negative similarity S_{ij} ($y_j \neq y_i$) is[1]:

$$\begin{aligned}\frac{\partial L_i}{\partial S_{ij}} &= -\frac{1}{|P(i)|}\frac{\partial\sum_{p\in P(i)} log\frac{exp(S_{ip}/\tau)}{\sum_{k\in V(i)\cup P(i)} exp(S_{ik}/\tau)}}{\partial S_{ij}} \\ &= -\frac{\sum_{p\in P(i)}}{|P(i)|}(\frac{\partial(S_{ip}/\tau)}{\partial S_{ij}} - \frac{\partial(log\sum_{k\in V(i)\cup P(i)} exp(S_{ik}/\tau)}{\partial S_{ij}}) \\ &= \frac{exp(S_{ij}/\tau)}{\sum_{k\in V(i)\cup P(i)} exp(S_{ik}/\tau)}\cdot\frac{1}{\tau}\end{aligned} \tag{2}$$

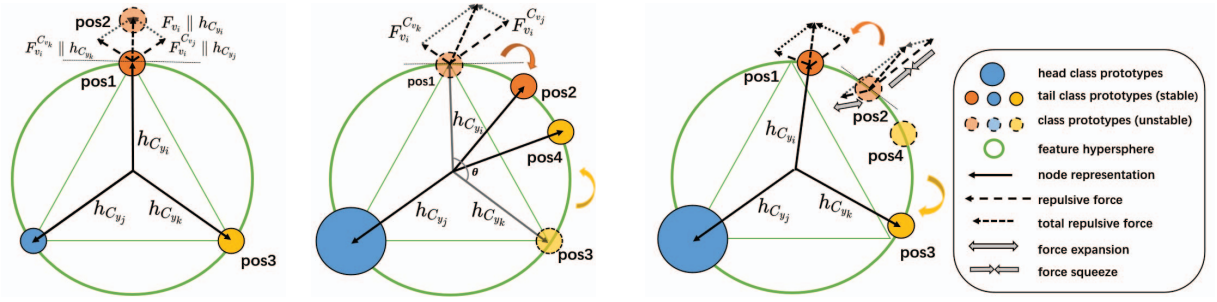
As node v_i and v_j belong to different classes, they are the negative sample of each other. Thus, S_{ij} can be updated by the gradient backward of the loss L_i and L_j. Similarly, we can compute the gradient of L_j with respect to S_{ji}. Let $\frac{\partial L_i}{\partial S_{ij}} = \alpha_j^i$ and $\frac{\partial L_j}{\partial S_{ji}} = \alpha_i^j$, where $S_{ij} = S_{ji}$. The gradient of the representation $\overrightarrow{h_i}$ provided by all negative nodes $\{v_j\}_{j=1}^{|V(i)|}$ can be explained as the repulsive force pushing the node v_i away from negative nodes in the embedding space. Specifically, given the total loss $L = \sum_{i\in V} L_i$, the gradient of L w.r.t. $\overrightarrow{h_i}$ provided by v_j can be formulated as:

$$\begin{aligned}\overrightarrow{F_{v_i}^{v_j}} &= -\frac{\partial L}{\partial S_{ij}}\cdot\frac{\partial S_{ij}}{\partial\overrightarrow{h_i}} = \frac{\partial L}{\partial S_{ij}}\cdot(-\overrightarrow{h_j}) = (\frac{\partial L_i}{\partial S_{ij}} + \frac{\partial L_j}{\partial S_{ji}})\cdot(-\overrightarrow{h_j}) \\ &= (\alpha_j^i + \alpha_i^j)\cdot(-\overrightarrow{h_j}),\end{aligned} \tag{3}$$

which can be thought of as analogous to a force altering v_i position that aims to push node v_i far away from node v_j. Then, given a class $C_{y_j} \in V(i), y_j \neq y_i$, we formulate the cumulative repulsive force imposed on v_i by the class C_{y_j} as:

$$\overrightarrow{F_{v_i}^{C_{y_j}}} = \sum_{x_m\in C_{y_j}}(\alpha_m^i + \alpha_i^m)\cdot(-\overrightarrow{h_m}) \approx N_{C_{y_j}}\cdot(\alpha_{C_{y_j}}^i + \alpha_i^{C_{y_j}})\cdot(-\overrightarrow{h_{C_{y_j}}}) \tag{4}$$

[1]As v_i and v_j form a pair of negative, $v_j \notin P(i)$. Therefore, $\frac{\partial S_{ip}}{\partial S_{ij}} = 0$.



(a) SCL on balanced classes (b) SCL on imbalanced classes (c) Our method on imbalanced classes

Fig. 2. Prototypes' distribution of (a) SCL with a balanced dataset when the loss converges. (b) SCL with an imbalanced dataset during training. (c) Our method with an imbalanced dataset during training. Node v_i, v_j, v_k belong to the orange class, blue class, and yellow class, respectively.

The last approximate equation in (4) holds due to the mild assumption that SCL leads to compact classes [16]. $N_{C_{y_j}}$ is the number of all nodes in class C_{y_i} and $\alpha^i_{C_{y_j}}$ denotes the approximated gradient of the loss w.r.t. C_{y_j}.

Note that it's common knowledge that CL/SCL conducts the representation learning on the unit hypersphere [17]. Therefore, in a scenario of balanced data, which has N classes with equal size, it resembles the Thomson Problem, which strives to find the equilibrium distribution of N electrons, which repels each other, constrained to the surface of a unit sphere.

Remark 1: The solution to Thomson Problem – the determination of the stable equilibrium configurations of N electrons confined to the surface of a sphere and repelling each other by a specified force law – corresponds to the representation assignment via gradient update w.r.t the SCL loss under a scenario of balanced data.

Suppose that there are 3 classes in a classification task. The stable equilibrium configurations of 3 classes are that the 3 electrons produce an equilateral triangle on the sphere, as depicted in Fig. 2(a). At this stage, the gradient in (2) has not been minimized to 0 as it requires an infinitely large distance between the particles. And it can never reach 0 because of the unit hypersphere requirement of CL/SCL. However, as we prove below, the gradient updates will still converge despite the fact that the updating gradient isn't 0.

Considering the case of three classes in Fig. 2(a), $\overrightarrow{h}_{C_{y_i}}$, $\overrightarrow{h}_{C_{y_j}}$, and $\overrightarrow{h}_{C_{y_k}}$ represent the prototypes of three classes, respectively. Similar to (4), the cumulative repulsive force imposed on v_i by the class C_{y_k} is deducted as

$$\overrightarrow{F_{v_i}^{C_{y_k}}} \approx N_{C_{y_k}} \cdot (\alpha^i_{C_{y_k}} + \alpha_i^{C_{y_k}}) \cdot (-\overrightarrow{h_{C_{y_k}}}). \tag{5}$$

Due to their similarity, it's reasonable to analogy the gradients $\overrightarrow{F_{v_i}^{C_{y_j}}}$ and $\overrightarrow{F_{v_i}^{C_{y_k}}}$ to the repulsive forces of two electrons in the Thomson Problem. Fig. 2(a) illustrates the stable convergence state. Based on (4) and (5), the information of the forces can be computed as below.

- **The magnitude of the force.** $|N_{C_{y_j}} \cdot (\alpha^i_{C_{y_j}} + \alpha_i^{C_{y_j}})|$ and $|N_{C_{y_k}} \cdot (\alpha^i_{C_{y_k}} + \alpha_i^{C_{y_k}})|$ mainly determine the magnitude of two forces, respectively, as $||\overrightarrow{h_{C_{y_j}}}|| = 1$ and $||\overrightarrow{h_{C_{y_k}}}|| = 1$.
- **The direction of the force.** Addressing the negative sign in front of the direction vector in (4)–(5), the directions of two forces $\overrightarrow{F_{v_i}^{C_{y_i}}}$ and $\overrightarrow{F_{v_i}^{C_{y_k}}}$ should be opposite to the directions of $\overrightarrow{h_{C_{y_j}}}$ and $\overrightarrow{h_{C_{y_k}}}$, respectively.

The gradients used to update v_i are composed of two parts, which are equivalent to the fact that the resultant force $\overrightarrow{F_{v_i}}$ is the sum of the repulsive forces $\overrightarrow{F_{v_i}^{C_{y_j}}}$ and $\overrightarrow{F_{v_i}^{C_{y_k}}}$ from the other two classes, given as:

$$\overrightarrow{F_{v_i}} = \overrightarrow{F_{v_i}^{C_{y_j}}} + \overrightarrow{F_{v_i}^{C_{y_k}}} \tag{6}$$

According to the symmetrical distribution of the hypothesized electrons, the resultant force $\overrightarrow{F_{v_i}}$ is perpendicular to the tangent line at the point of v_i[2]. According to (4)–(6), the resultant force given to v_i is parallel to the node embedding $\overrightarrow{h_i}$: $\overrightarrow{F_{v_i}} \parallel \overrightarrow{h_i}$. During training, the stochastic gradient descent optimizer will combine forces, i.e., gradients and learning rate l, to optimize the loss function:

$$\overrightarrow{\widetilde{h}_i^{(t+1)}} = \overrightarrow{h_i^{(t)}} - l \cdot \frac{\partial L}{\partial h_i}\bigg|_{h_i = h_i^{(t)}} = (1+\lambda) \cdot \overrightarrow{h_i^{(t)}} \tag{7}$$

$\overrightarrow{h_i^{(t)}}$ denotes the representation of node v_i at epoch t. It should satisfy the stable equilibrium configurations of Thomson problem, and $\lambda = l * ||\overrightarrow{F_{v_i}}|| > 0$. Consequently, the node embedding h_i moves from pos1 to pos2 (Fig.2(a)). As CL requires that the latent representations are limited to the unit hypersphere through normalization we then have :

$$\overrightarrow{h_i^{(t+1)}} = \frac{\overrightarrow{\widetilde{h}_i^{(t+1)}}}{||\overrightarrow{\widetilde{h}_i^{(t+1)}}||} = \overrightarrow{h_i^{(t)}}. \tag{8}$$

Thus, we can conclude that gradient-based updates driven by SCL are now "converged".

Remark 2: Even if the gradient isn't zero, showing that the loss hasn't reached the global minimum, the model has reached a stable state where the embedding no longer updates with epochs. Perfect uniformity among classes indicates that they are distinguishable in the latent embedding space.

[2]All nodes of the same class are considered to be located at the same spot during classes-level analysis.

V. Gradient Analysis for Imbalanced Data

Though CL/SCL works flawlessly for balanced datasets, however, this is no longer the case for imbalanced datasets. Assume that many samples are added to the blue class, making it the head class while the others remain unchanged. The blue electron would now have greater charge than the others. Therefore, for node v_i, the repulsive force from class C_{y_j} is much larger than that from class C_{y_k}. Consequently, the resultant force $\overrightarrow{F_{v_i}}$ would push v_i to rotate clockwise during the backpropagation. Similarly, v_k would rotate counterclockwise. Eventually, the model reaches a stable state when the orange and yellow particles move to pos2 and pos4 (Fig. 2(b)), where the components of the tangential directions of $\overrightarrow{F_{v_{i/k}}^{C_{y_{j/i}}}}$ and $\overrightarrow{F_{v_{i/k}}^{C_{y_{k/j}}}}$ at pos2 (Fig. 2(b)) cancel each other.

Note that the orange and yellow particles are equally charged, which is consistent with a mild assumption that the sizes of the tail classes are comparable, considering the head classes typically dominate the entire dataset. It is reasonable to infer that the three particles constrained in the hypersphere form an isosceles triangle, as shown in Fig. 2(b). Denoting $\angle(\overrightarrow{h_{C_{y_i}}}, \overrightarrow{h_{C_{y_k}}})$ as θ, the deduction with the fundamental electromagnetic and geometric principles leads to:

$$\frac{||\overrightarrow{F_{v_i}^{C_{y_j}}}||}{||\overrightarrow{F_{v_i}^{C_{y_k}}}||} = \frac{cos(\theta - \pi/2)}{sin(\theta/2)} = \frac{sin(\theta)}{sin(\theta/2)} = 2cos(\theta/2) \tag{9}$$

where $\theta = arccos(S_{ik})$. As $||\overrightarrow{F_{v_i}^{C_{y_j}}}||$ is greater than $||\overrightarrow{F_{v_i}^{C_{y_k}}}||$ in the imbalanced setting, θ is smaller than that in Fig. 2(a).

Remark 3: The uniformity pursued by SCL no longer holds in imbalanced scenarios. The dominant head classes could drive the tail classes indistinguishable.

Sharpen the repulsive force for imbalanced data

As the imbalanced data compromises the embedding uniformity, the representations of tail classes will intertwine. To address the issue, recalibration should be applied to the repulsive force to close the gap of sample numbers between classes. Specifically, the repulsive force should increase substantially when prototypes of different classes are too close and decrease substantially when they are too far away. Thus, we propose to rescale the repulsive force by applying linear scaling to the gradient. Then, the corresponding GR-SCL Loss becomes:

$$L_i^{GR-SCL} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(S_{ip}/\tau)}{\sum_{k \in V(i) \cup P(i)} exp(\gamma \cdot S_{ik}/\tau)} \tag{10}$$

The magnitude of the repulsive force between v_i and v_j is

$$\begin{aligned} ||\overrightarrow{F_{v_i}^{v_j}}|| &= \frac{\partial L_i^{GR-SCL}}{\partial S_{ij}} = \frac{\gamma}{\tau} \cdot \frac{exp(\gamma \cdot S_{ij}/\tau)}{\sum_{k \in V(i) \cup P(i)} exp(\gamma \cdot S_{ik}/\tau)} \\ &= \frac{\gamma}{\tau} \cdot \frac{exp(S_{ij}/\tau)}{\sum_{k \in V(i) \cup P(i)} \frac{exp(S_{ik})}{\tau} \cdot \underbrace{exp(\frac{(\gamma - 1)}{\tau} \cdot (S_{ik} - S_{ij}))}_{contraction\ factor}} \end{aligned} \tag{11}$$

where γ is the scaling factor that controls the sharpness of the force/gradient. The proposed solution is a generalized form of SCL and is specialized to deal with imbalanced data. When $\gamma = 1$, the loss degenerates to the conventional SCL loss.

Suppose node v_j is the most distant negative sample for v_i ($\forall v_k \in V, S_{ik} \geq S_{ij}$), when $\gamma > 1$, the contraction factor becomes no smaller than 1, and the greater γ, the greater the contraction factor. In comparison with (2), the gradient of the loss with respect to S_{ij} is smaller, which mitigates the dominating impact from head classes by their large numbers of samples. Similarly, our solution increases the gradient for the hardest sample to counteract the impact of insufficient samples of tail classes. This allows the gradient-based updates to converge before class C_{y_i} reaches pos2 as shown in Fig. 2(c), leaving a clearer boundary between the two tail classes.

Although based on a 3-class configuration, the GR-SCL can be easily extended to multi-class configurations.

VI. Experiment

We extensively experimented, including imbalanced node classification and ablation study, on four datasets to indicate the efficacy of the proposed method, aiming to answer the following questions: $\mathcal{Q}1$: How effective is GR-SCL Loss compared with the SOTA baselines on imbalanced node classification tasks under different imbalance ratios? $\mathcal{Q}2$: Can GR-SCL Loss learn the node embeddings such that the representation of each class achieves better uniformity in the hypersphere?

A. Experimental setup

Datasets For a comprehensive comparison, we experimented on four widely-used datasets for imbalanced node classification tasks, including Amazon-computers, Amazon-photo [18], CoraFull, and Cora-ML [19]. To eliminate the impact of imbalanced data during evaluation, the sample size for each class in the test set is the same. The training set was sampled with an exponential decay across classes, controlled by the imbalance ratio ρ. We set ρ to 20, 50, 100, and 200.

Comparison Models

- **GCN[20]+RW**: A reweighting method that gives higher weight to tail classes and lower weight to head classes.
- **GraphSMOTE** [8]: An interpolation-based method that synthesizes minority samples in the embedding space.
- **ImGAGN** [10]: A generative adversarial graph network model which utilizes a generator to produce synthetic minority nodes and a discriminator to discriminate between minority nodes and majority nodes.
- **DPGNN** [11]: A class prototype-driven method uses distance metric learning to balance the sample difference.
- **OURS**: Following [21], we employed a two-stage strategy to implement our method. In the 1st stage, we generated graph views by masking node features and removing edges, and then applied GR-SCL Loss. In the 2nd stage, a linear classifier was trained on top of the learned representation with loss re-weighting.
- **OURS[-scl][w/o RW]**: Variants of OURS, with "-scl" indicating that GR-SCL loss is replaced with the SCL loss and "w/o RW" indicating that the reweight strategy for the classifier training is spared for the 2nd stage.

Evaluation protocol The performance is measured with Top1-accuracy over the entire dataset. Following [21], the classes

Dataset	Amazon-computers				Amazon-photos				CoraFull				Cora-ML			
Imbalance ratio	20	50	100	200	20	50	100	200	20	50	100	200	20	50	100	200
ImGAGN	0.735	0.696	0.644	0.606	0.791	0.752	0.694	0.646	0.441	0.436	0.420	0.387	0.681	0.614	0.578	0.53
$GCN + RW$	0.903	0.879	0.849	0.793	0.922	0.880	0.8265	0.806	0.575	0.576	0.554	0.517	0.848	0.803	0.789	0.616
GraphSMOTE	0.894	0.805	0.787	0.751	0.925	0.895	0.879	0.833	0.526	0.521	0.485	0.445	0.83	0.76	0.67	0.563
DPGNN	0.903	0.893	0.889	0.897	0.9325	0.9205	0.9185	0.912	0.355	0.323	0.319	0.302	0.862	0.823	0.806	0.721
OURS-scl	0.912	0.908	0.903	0.901	0.928	0.928	0.918	0.915	0.554	0.586	0.577	0.532	0.843	0.807	0.771	0.697
OURS	**0.934**	**0.933**	**0.929**	**0.922**	**0.946**	**0.938**	**0.932**	**0.925**	**0.592**	**0.611**	**0.601**	**0.558**	**0.882**	**0.844**	**0.822**	**0.757**

TABLE I: The imbalanced node classification results on four real-world datasets. The best performance is marked in bold.

were divided into three subsets, including *head, medium, tail*, based on the number of samples in each class, and the average subset accuracy of which was also recorded.

B. Imbalanced Classification Performance (Q1)

Table I reports the node classification performance of all methods. It's evident that the proposed GR-SCL Loss substantially outperforms all baselines in all datasets with varying and significantly large imbalance ratios, proving the effectiveness of our method. We can also observe that the performance of ImGAGN, GCN+RW, and GraphSMOTE decreases significantly with the increase of the imbalance ratio ρ, indicating that they fail to differentiate samples from tail classes. Notably, DPGNN shows rather inferior performance on CoraFull. CoraFull has the largest number of classes among all datasets. As DPGNN's distance metric learning module requires concatenating the difference between the node's embedding and all class prototypes, the dimension of the concatenated embedding for CoraFull would be quite large, resulting in the curse of high dimensionality [12]. In comparison, our model is robust to the highly imbalanced dataset.

C. Ablation Study (Q2)

Fig. 3 presents the performance of the variants of our model. No matter if the reweighting strategy is utilized, it is evident that our proposed GR-SCL Loss contributes prominently to performance improvement.

Table II lists the performance of our model on *Head*, *Medium*, *Tail*, and *All* subsets/classes with different imbalance ratios. Specifically, we excluded the reweighting strategy/component to reflect the effectiveness of GR-SCL Loss exclusively. It's clear that *Head* and *Medium* classes have slight performance enhancement, whereas *Tail* classes have significant enhancement, demonstrating that our method can discover the representation for each class with a distinct decision boundary, especially for the tail classes.

ρ	methods	Head	Medium	Tail	**All**
20	OURS w/o RW	0.923	0.965	0.863	0.912
	OURS-scl w/o RW	0.916	0.958	0.787	0.877
50	OURS w/o RW	0.921	0.961	0.821	0.893
	OURS-scl w/o RW	0.916	0.951	0.746	0.859
100	OURS w/o RW	0.926	0.968	0.766	0.875
	OURS-scl w/o RW	0.924	0.94	0.713	0.845
200	OURS w/o RW	0.93	0.960	0.576	0.797
	OURS-scl w/o RW	0.928	0.936	0.527	0.770

TABLE II: Detailed performance of *Head*, *Medium*, *Tail*, and *All* classes on OURS and OURS-scl with varying ρ.

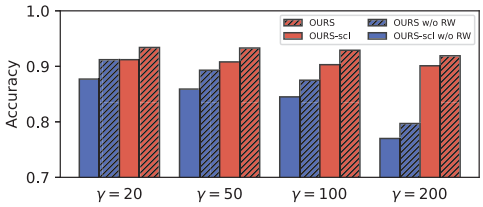


Fig. 3. Ablation study on Amazon-computers with different ρ.

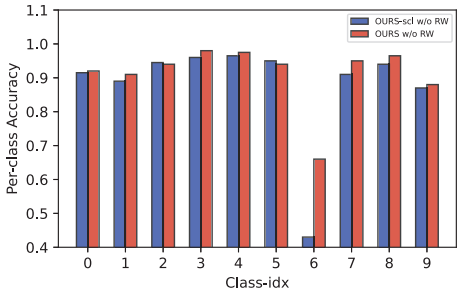
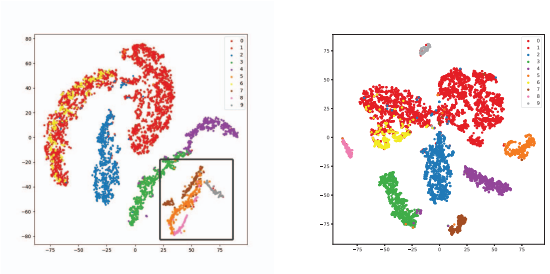


Fig. 4. Class performance between OURS w/o RW and OURS-scl w/o RW on Amazon-Computers with $\rho = 20$.



(a) OURS-scl (b) OURS

Fig. 5. The feature visualization with t-SNE on Amazon-Computers.

For more detailed insight, Fig. 1, 4, and 5 illustrate the qualitative and quantitative per-class analysis of OURS from different perspectives. Fig. 1(b) shows that some classes, such as 7, 8, and 9, share high similarities, indicating that tail

classes are intertwined in the representation space, as depicted in Fig 5(a). In comparison, in Fig 1(c), the heat map of node similarity matrices in the representation space tends to be a matrix consisting of all-ones matrices on the diagonal, which indicates that after the gradient recalibration, the representation entanglement between classes is substantially eased, showing better uniformity as depicted in Fig 5(b). Relatively orthogonal features for each class allow a classifier to better learn explicit classification boundaries. Therefore, it is not surprising to see these classes achieve better performance than the conventional SCL-based approach, as shown in Fig. 4. For example, tail class-6 is intertwined with class-1, which has a dominating number of samples so as to prevent correct identification of class-6, resulting in degraded performance. In contrast, our model helps to significantly improve the visibility and classification performance of tail classes.

VII. Conclusion

In this paper, we mainly focused on the imbalanced node classification problem. This problem significantly influences the performance of tail classes in supervised learning. To quantitatively model the impact of class sizes on representation learning, we used repellent electrons on the surface of a unit hyper-sphere to simulate the search for equilibrium states of different class configurations. We applied the stable states in the Thomson Problem as the solutions for balanced classes, and analyzed the reason for representation collapse in tail classes from the perspective of the gradient. To alleviate this problem, we recalibrated the gradient to close the gap among classes, leading to a more balanced and uniform feature space. Experiments on multiple imbalanced graph datasets demonstrate the effectiveness of our method, which outperforms the other baselines. The ablation study and case study also verify the representation collapse of tail classes is mitigated.

VIII. acknowledge

This work was partially supported by the NSFC under Grants 92270125, 62276024, and 72074209, as well as U2336201.

References

[1] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *NeurIPS*, 2020.

[2] S. Pan and X. Zhu, "Graph classification with imbalanced class distributions and noise," in *IJCAI*, Citeseer, 2013.

[3] Z. Zhou, J. Yao, F. Hong, Y. Zhang, B. Han, and Y. Wang, "Combating representation learning disparity with geometric harmonization," in *NeurIPS*, 2023.

[4] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Deep graph contrastive representation learning," 2020.

[5] J. J. Thomson, "Xxiv. on the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 7, no. 39, pp. 237–265, 1904.

[6] S. Ando and C. Y. Huang, "Deep over-sampling framework for classifying imbalanced data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 770–785, Springer, 2017.

[7] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," 2019.

[8] T. Zhao, X. Zhang, and S. Wang, "Graphsmote: Imbalanced node classification on graphs with graph neural networks," in *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 833–841, 2021.

[9] M. Shi, Y. Tang, X. Zhu, D. Wilson, and J. Liu, "Multi-class imbalanced graph convolutional network learning," in *IJCAI*, 2020.

[10] L. Qu, H. Zhu, R. Zheng, Y. Shi, and H. Yin, "Imgagn: Imbalanced network embedding via generative adversarial graph networks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1390–1398, 2021.

[11] Y. Wang, C. Aggarwal, and T. Derr, "Distance-wise prototypical graph neural network in node imbalance classification," 2021.

[12] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*, pp. 420–434, Springer, 2001.

[13] F. Hong, J. Yao, Z. Zhou, Y. Zhang, and Y. Wang, "Long-tailed partial label learning via dynamic rebalancing," in *ICLR*, OpenReview.net, 2023.

[14] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[15] F. Hong, J. Yao, Y. Lyu, Z. Zhou, I. Tsang, Y. Zhang, and Y. Wang, "On harmonizing implicit subpopulations," in *ICLR*, OpenReview.net, 2024.

[16] F. Graf, C. Hofer, M. Niethammer, and R. Kwitt, "Dissecting supervised constrastive learning," in *ICML*, pp. 3821–3830, PMLR, 2021.

[17] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *ICML*, pp. 9929–9939, PMLR, 2020.

[18] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann, "Pitfalls of graph neural network evaluation," 2018.

[19] A. Bojchevski and S. Günnemann, "Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking," 2017.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016.

[21] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for representation learning," in *ICLR*, 2020.