Contrastive Learning Meets Homophily: Two Birds with One Stone

Dongxiao He [1] Jitao Zhao [1] Rui Guo [1] Zhiyong Feng [1] Di Jin [1] Yuxiao Huang [2] Zhen Wang [3] Weixiong Zhang [4]

Abstract

Graph Contrastive Learning (GCL) has recently enjoyed great success as an efficient self-supervised representation learning approach. However, the existing methods have focused on designing of contrastive modes and used data augmentation with a rigid and inefficient one-to-one sampling strategy. We adopted node neighborhoods to extend positive samplings and made avoided resorting to data augmentation to create different views. We also considered the homophily problem in Graph Neural Networks (GNNs) between the inter-class node pairs. The key novelty of our method hinged upon analyzing this GNNs problem and integrating the GCL sampling strategy with homophily discrimination, where we solved these two significant problems using one approach. We introduced a new parameterized neighbor sampling component to replace the conventional sub-optimal samplings. By keeping and updating the neighbor sets, both the positive sampling of GCL and the message passing of GNNs can be optimized. Moreover, we theoretically proved that the new method provided a lower bound of mutual information for unsupervised semantic learning, and it can also keep the lower bound with downstream tasks. In essence, our method is a new self-supervised approach, which we refer to as group discrimination, and it can make the downstream fine-tuning efficient. Our extensive empirical results demonstrate that the new method can significantly outperform the existing GCL methods because the former can solve the homophily problem in a self-supervised way with the new group discrimination method used.

1. Introduction

Graph representation learning aims at learning characteristic low-dimensional representations for network structured data to facilitate diverse downstream tasks (Zhu et al., 2019; Cai et al., 2018). It has many applications in various fields, such as biology (Duvenaud et al., 2015), social networks (Fan et al., 2019), finance (Wu et al., 2019), and many text-rich graph scenarios (Yu et al., 2021b). The current methods of deep learning on graphs take advantage of the expressive power of neural networks to encode graph patterns (Wu et al., 2021; Scarselli et al., 2009). However, these methods require graph labels, which are challenging and require domain knowledge, making supervised end-to-end training difficult for GNNs. Inspired by the success of self-supervised learning in other fields (Jing & Tian, 2021; Schmarje et al., 2020), graph contrastive learning has made remarkable progress that is comparable to supervised learning but requires no labeled data (Liu et al., 2022; Zhu et al., 2021b). The general objective of contrastive learning is to make positive pairs of instances closer in the embedding space and to keep negative pairs farther apart, where every two instances in a positive pair are supposed to have the same label in downstream tasks. It achieves so mostly by maximizing the similarity between two data augmentation views of the same graphs/nodes (*a.k.a.* positive samplings) and minimizing the similarity between different graphs/nodes (*a.k.a.* negative samplings).

However, the one-to-one sampling strategy of the existing GCL methods (Zhu et al., 2020b; Thakoor et al., 2021; Bielak et al., 2021) follow the fixed paradigm limited to the same index between augmented views. It builds upon the label consistency assumption (Wang et al., 2022) that data augmentations would not change the semantic information. However, this is too strict for GCL compared to visual contrastive learning. First, the semantic labels of graphs are abstract compared to that of images, which largely depend on domain knowledge and specific conditions. The existing data augmentation methods based on prior knowledge or random uniform distribution may potentially alter node/graph semantics. Second, representation learning

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China [2]Department of Data Science, George Washington University, NW Washington DC, America [3]Department of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China [4]Department of Health Technology and Informatics, Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China. Correspondence to: Rui Guo <961941663@qq.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

on graphs depends on graph topologies and attributes because of the unique message-passing encoding mechanism of GNNs (Kipf & Welling, 2017). However, the existing data augmentation methods all adopt some types of perturbations to the structure and features of the graph, which will affect the representation learning of graph data (Zhu et al., 2020b; Thakoor et al., 2021). In short, traditional sampling strategies inherited from visual learning extensively rely on data augmentation. This could be problematic for GCL since label consistency cannot be guaranteed. We were motivated to develop a new contrastive learning approach that significantly deviated from the existing methods by using no data augmentation.

In GCL, pairs of nodes belonging to the same class are considered as positive samplings. Interestingly, the most fundamental homophily problem of GNNs also faces the same issue where finding other intra-class nodes for the anchor node is required. GNNs work under the homophily assumption (Zhu et al., 2020a), i.e., most connected nodes are from the same class and have similar features. However, message passing on non-homophily edges leads to fusing of information from different classes. It degrades the downstream tasks' performance (Zheng et al., 2022), result in different designs for heterophily situations (Jin et al., 2021a). Therefore, if it is possible to identify edges linking intra-class node pairs during message passing, which is the same as the most important criterion for designing positive samplings in GCL, it will reduce the noise propagation of non-homophily edges and improve the performance of downstream tasks.

Therefore, two questions arise naturally, i.e., *How to integrate the sampling problem of graph contrastive learning and the homophily problem of graph neural networks, and how to solve them simultaneously?* The previous GCL works developed various data augmentation methods, e.g., based on parameterization (Suresh et al., 2021; Kefato et al., 2021b; You et al., 2021), prior data distributions (Xia et al., 2022), or additional domain knowledge (Hassani & Ahmadi, 2020; Jin et al., 2021b). However, they all created a second view for the one-to-one sampling strategy, and various prior-based augmentation schemes they used were typically task-related, compromising the performance of model generalization. Methods for addressing the homophily problem of GNNs usually seek alternative ways of message passing, such as weighted neighbor and higher-order neighbor propagation. But they all ignore to address the most fundamental cause of performance degradation, i.e., the existence of non-homophily edges. Thus, they were typically sub-optimal for solving this homophily problem.

In this paper, we approached these questions by introducing a new graph contrastive learning method based on the neighbor sampling and homophily discrimination, namely **Ne**ighbor **Co**ntrastive Representation Learning on Graphs (**NeCo**). We creatively propose to expand the scope of GCL samplings to neighbor sets for anchor nodes. Therefore, we can abandon the inherent two-view settings in GCL models, and as a result, graph data augmentation is no longer required. Considering the homophily problem, that graph structure with weak homophily may have a negative impact on the proposed NeCo and GNNs encoders, we introduce a GCL model with additional parameterized homophily discrimination modules. It iteratively updates and maintains the neighbor sets using neural networks and sampling tricks at each epoch, which optimize the encoding of GNNs as well as the positive samplings of GCL simultaneously. Moreover, we theoretically prove that, with the homophily improvement to the learned graph structure, our parametric neighbor sampling strategy can guarantee a lower bound of mutual information on encoding semantic representation and learning downstream task-specific representation. More importantly, NeCo provides a new pretext task for self-supervised learning, which we call group discrimination. The new task can reduce the difficulty of fine-tuning on the downstream task networks.

Finally, we experimentally demonstrate the performance of our NeCo method on node classification tasks using five heterophily datasets and four commonly used homophily datasets. The NeCo method significantly improves performance compared to the state-of-the-art baselines. Furthermore, the learned topologies with stronger homophily significantly improve the classical GNNs. Therefore, NeCo can also be regarded as a novel idea for solving the inherent homophily problem of GNNs.

2. Problem Definition

We consider an undirected graph $G = (\mathcal{V}, \mathcal{E})$ with a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$, and use X to denote a feature matrix. The dependency between two random variables B and C can be measured by the mutual information, written as $I(B; C)$.

2.1. Graph Contrastive Learning

GCL methods learn encoders by maximizing the mutual information between the representations of graph G (or node v_i) and its positive samples, which can be expressed as $I(f(G); f(G_+))$. It is equivalent to $I(G; f(G))$ (as known as the InfoMax principle), which aims at making the encoded features of graph G easily distinguished from others in the representation space.

The positive samples of the existing GCL methods are typically generated by graph data augmentation (GDA). Based on the assumption that GDAs would not change semantics labels, multiple views of the same graph are regarded as

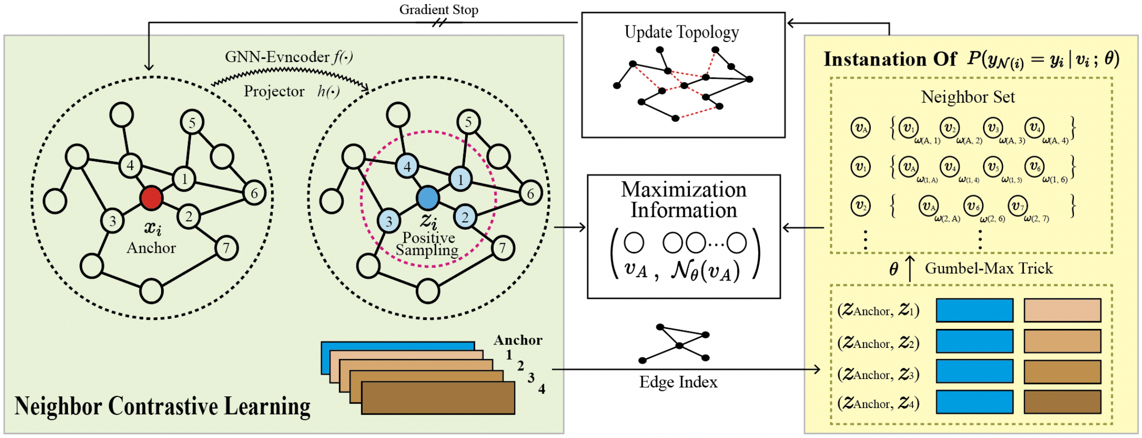

Gradient Stop
Update Topology
GNN-Evncoder $f(\cdot)$
Projector $h(\cdot)$
5
4
1
6
3
2
7
x_i
Anchor
z_i
Positive
Sampling
Maximization
Information
v_A , $\mathcal{N}_\theta(v_A)$
Instanation Of $P(y_{\mathcal{N}(i)} = y_i \mid v_i\,;\theta)$
Neighbor Set
v_A $\{v_1\ \omega_{(A,1)}\ v_2\ \omega_{(A,2)}\ v_3\ \omega_{(A,3)}\ v_4\ \omega_{(A,4)}\}$
v_1 $\{v_A\ \omega_{(1,A)}\ v_4\ \omega_{(1,4)}\ v_5\ \omega_{(1,5)}\ v_6\ \omega_{(1,6)}\}$
v_2 $\{v_A\ \omega_{(2,A)}\ v_6\ \omega_{(2,6)}\ v_7\ \omega_{(2,7)}\}$
θ Gumbel-Max Trick
(z_{Anchor}, z_1)
(z_{Anchor}, z_2)
(z_{Anchor}, z_3)
(z_{Anchor}, z_4)
Anchor
1
2
3
4
Edge Index
Neighbor Contrastive Learning

Figure 1. An overview of our proposed method NeCo for graph representation learning and homophily discrimination. It contains two main components: The GNN encoder extracts features and collects positive samplings from neighbors for GCL. The homophily discrimination module keeps and updates the learned neighbor sets.

positive samples for each other, and correspondingly, different graphs are regarded as negative samples. InfoNCE is a commonly used mutual information estimator implemented by making the cosine similarity between pairs of positive samples greater than that of negative samples. The objective of GCL methods is as follows:

$$l_n = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp(sim(z_i, z_{i'})/\tau)}{\sum_{j=1, j\neq i'}^{2n-1} \exp(sim(z_i, z_j)/\tau)} \tag{1}$$

2.2. Homophily

We focus on the node-level representation learning for message passing of GNNs. For a node $v_i \in \mathcal{V}$ in the graph G, with the corresponding representation z_i that initialized as the input features in X, we define the $\mathcal{N}$ as the set of nodes directly connected to the anchor v_i, that is $\mathcal{N}(i) = \{v_j | (v_i, v_j) \in \mathcal{E}\}$. Then the unified message passing method that most GNNs use can be written as:

$$z_i^{(k)} = \text{UPDATE}^{(k)}\Big(z_i^{(k-1)}, \\ \text{AGGREGATE}^{(k)}\left(\{z_j^{(k-1)} \mid j \in \mathcal{N}(i)\}\right)\Big) \tag{2}$$

where AGGREGATE($\cdot$) and UPDATE($\cdot$) are trainable functions used for neighbor propagation and representation update of node v_i, respectively. We also consider both homophily and heterophily in class labels. To evaluate the homophily level of graph structure, we define the homophily ratio as:

Definition 1 (Homophily Rate h & β). *For a graph G, the **label homophily rate** h measures the fraction of edges with intra-class node pairs. It can be interpreted as the probability $P(y_j = y_i, j \in \mathcal{N}(i) \mid v = v_i)$. The **attribute homophily rate** β measures the proportion of consistency attribute dimensions between neighboring nodes. They can be implemented by*

$$h = \frac{\sum_i |\mathcal{Y}(\mathcal{N}(i)) \cap \mathcal{Y}(i)|}{2|\mathcal{E}|}, \quad \beta = \frac{\sum_i \sum_{j\in\mathcal{N}(i)} (X_i \odot X_j)}{2|\mathcal{E}|} \tag{3}$$

Where $\mathcal{N}(i)$ means a set composed of the neighbors of node i as mentioned earlier, $\odot$ denotes dividing the number of dimensions with the same value between two vectors by the number of all dimensions. For simplicity, we will denote the probability above as $P(y_{\mathcal{N}(i)} = y_i \mid v_i)$ hereafter. Datasets with higher h and β are referred to as homophily graphs. Weaker homophily datasets with lower values are referred to as heterophily graphs.

3. Methods

3.1. Motivation

We start with the question of *How the homophily/heterophily node pairs affect GNNs and GCL?* We look for answers from the motivation experiments to show that GNNs and GCL may face the same homophily (or the intra-class positive sampling) problem, and both of them will benefit from solving it. For GNNs, we randomly drop a certain rate t of heterophily edges for message passing. Correspondingly, we extend the samplings of GCL methods to the neighbor sets $\mathcal{N}(i)$, and use the same t to control the number of inter-class positive samplings. Figure 2 shows the results. As the graph's rate t for node pairs increases, both GNNs and GCL methods significantly improve upon the traditional GCN and GRACE with the original topology. This observation in-

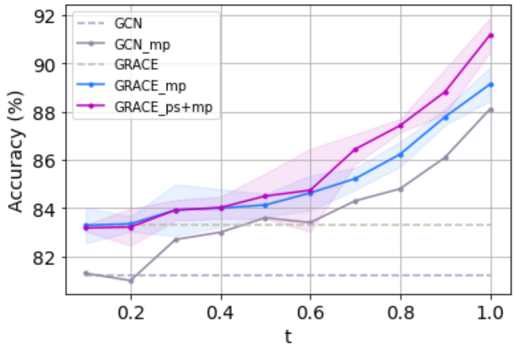


Figure 2. The homophily of node pairs affects GNNs and GCL to a large extent. t is the ratio of removed heterophily edges. "mp" and "ps" means that the optimized topology participates in **m**essage **p**assing and the **p**ositive **s**amplings, respectively.

dicates that the neighbor set $\mathcal{N}(i)$ is a rich source of positive samples that GCL has always overlooked. Besides, it indicates that homophily discrimination is the key to improving both GNNs and GCL models.

Two requirements that an effective sampling strategy for contrastive learning needs to meet. The first is label consistency (Wang et al., 2022), meaning that labels between positive samples must be consistent for learning meaningful representations. The second is the information gap (Xu et al., 2021; Wei et al., 2021), meaning that based on the guarantee of label consistency and information bottleneck theory, a large attribute gap between positive samplings would improve the performance of contrastive learning. The existing GCL methods all attempt to satisfy these two requirements. It is similar to the idea of solving the problem of homophily that propagates robust information by detecting edges that satisfy label consistency but in an unsupervised way. The information defined by homophily rate h and β could be regarded as a robust feature required by information bottleneck theory and message passing. Then any perturbation to the graph structure or attributes that all GDAs adopt would inevitably distort useful information. We provide more details in the Appendix.

Therefore, the neighbor set $\mathcal{N}(i)$ is a natural, easy-to-get (compared to the GDAs) and efficient choice (as shown in the motivation experiment) for sampling in GCL methods. The selection of positive samplings in GCL can also be facilitated to solve the homophily problem of GNNs since contrastive learning is known for its strong ability to learn semantics without labels.

3.2. The NeCo Framework

We now introduce our NeCo approach based on the efficient parameterized neighboring sampling and homophily discrimination, as shown in Figure 1. Our design hinges upon using the powerful semantic learning ability of GCL to empower the capacity of homophily discrimination for message passing of GNNs, and the homophily structure can guide the positive samplings for GCL.

Neighboring Sampling As shown in the motivation experiments, positive samplings within the first-order neighbors can effectively avoid the problem in one-to-one samplings based on GDAs with superior performance. Then, we update the single positive sampling in Equation 1 to be the sum of the similarity between the anchor node v_i and its neighbors in $\mathcal{N}(i)$. Since there is no second view, we treat other nodes $\{\mathcal{V} - \mathcal{N}(i)\}$ in the intra-view as negative samplings of the anchor v_i. Moreover, considering that discrete points in non-fully connected graphs may be problematic in no neighbor pairs to contrast, we add the self-loops to the positive sampling sets.

Parameterizing the Homophily Discrimination Note that the probability $P(y_{\mathcal{N}(i)} = y_i|v_i)$ shown in Section 2.2, which can discriminate homophily nodes, is vital in improving GCL and GNNs. We treat the homophily probability P to follow a Bernoulli distribution, and the random variable of P is implemented by connecting the parametric GCL sampling pairs directly after the encoder and projector. Inspired by the weight design of (Velickovic et al., 2018), we operate on the representations obtained by the final-layer of GCL model:

$$P(y_{\mathcal{N}(i)} = y_i|v_i;\theta) = \text{CONCAT}(h_i, h_{\mathcal{N}(i)})W_\theta \quad (4)$$

where the NeCo updates the homophily neighbor sets $\mathcal{N}(i;\theta)$ for each node in every iteration. Moreover, we employ the Gumbel-Max trick (Maddison et al., 2017) to turn this continuous probability of node pairs into categorical samplings, as $\arg\max\,(log(P(y_{\mathcal{N}(i)} = y_i|v_i;\theta) + \mathcal{G}_i))$, where $\mathcal{G}_i = -\log(-\log(U_i))$, $U_i \sim Uniform(0,1)$. Note that the learned stronger homophily structure is used in both the message passing in GNNs and positive samplings in GCL.

Overview As shown in Figure 1, in our NeCo model, we first generate node embeddings z_i by encoding with GNNs, and map the embeddings into the contrastive space $\mathcal{H} = \{h_i\}$. Then, we employ the parametric homophily discrimination module to compute the sampling probability $P(y_{\mathcal{N}(i)} = y_i|v_i;\theta)$. With the learned θ, we propose the neighboring pairwise loss based on homophily discrimination. We add a regularization term to guarantee the boundary of parametric neighbor samplings because the design of the loss function may have to sample all neighbors, making the model converge to a sub-optimal solution. After each iteration, we update the topology of the input graph with the result of homophily discrimination without backward pass of gradients. The final objective can be written as:

$$\arg\min_{\theta,f} -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\sum_{k\in\mathcal{N}(i;\theta)}\exp\left(sim(z_i,z_k)/\tau\right)}{\sum_{j\in\{\mathcal{V}-\mathcal{N}(i;\theta)\}}\exp\left(sim(z_i,z_j)/\tau\right)} + \lambda||\sum_{i\in\mathcal{V}}\mathcal{N}(i;\theta)||/|\mathcal{E}| \quad (5)$$

3.3. Theoretical Analysis

We propose the strategy for GCL positive samplings in the node neighbor sets. To our best knowledge, this is the first time to change the topology structure with no supervision for jointly considering graph contrastive learning. Thus, our discussion focuses on the relationship between the mutual information I and the neighbor sampling strategy $P(y_{\mathcal{N}(i)} = y_i|v_i;\theta)$, which can also be interpreted as the homophily rate of the learned structure h_θ. Furthermore, we analyze the capability that can bring to the GNNs for discriminating the heterophily node pairs. The new neighboring contrasting paradigm leads to new pretext tasks for self-supervised learning, which we call the intra-class group discrimination. It is superior to the instance discrimination task that previous GCL models adopt. The details of the proofs are provided in the Appendix.

First, based on the design of neighbor contrastive loss as Equation 5, we have,

Theorem 1. *Suppose that each node v_i has a homophily rate h_i, which could be defined as $h_i = |\mathcal{Y}(\mathcal{N}(i)) \cap \mathcal{Y}(i)|/degree_i$. Our method maximizes the lower bound of the mutual information between the representations of the contrastive pairs in NeCo, and the gap is determined by the learned homophily rate $h_{i,\theta}$ of the graph G. Specifically,*

$$-L \le I(f(v_i), f(\mathcal{N}(i;\theta))) + \mathbb{E}_{P(v_i)}(h_{i;\theta} - \frac{1}{|\mathcal{N}(i;\theta)|})\sum sim(z_i, z_{\mathcal{N}(i;\theta)})/\tau \quad (6)$$

This statement means that NeCo maximizes a lower bound of the mutual information between the representations of node v_i and its learned neighbors $\mathcal{N}(i;\theta)$. This bound indicates that, as the homophily rate of the structure discovered by NeCo increases (both for the whole graph with h_θ and the individual nodes with $h_{i,\theta}$), the mutual information lower bound of the neighboring sampling contrastive strategy will be improved.

Recently, some works (Suresh et al., 2021; Xu et al., 2021) discuss the information bottleneck theory that considers the mutual information between the input graphs G (or the embeddings z) and downstream task labels Y when designing models. To further explain that the information in neighbor sets can also achieve the same effect that captures certain overlapping information with the downstream task label of the anchor node, we get,

Theorem 2. *With the defined homophily rate h_i for each node, the mutual information between the representations and the downstream task labels Y can be written as:*

$$I(f(v_i);Y) \ge I(N(v_i;\theta);Y) = I(v_i;Y) + (h_{i,\theta}-1)(H(Y|v_i) - \sum_{y\in\mathcal{Y}}\frac{log(|\mathcal{Y}|-1)}{|\mathcal{Y}|-1}) \quad (7)$$

The first inequality in the theorem can be obtained from the data processing inequality as (Suresh et al., 2021) shows. It guarantees a lower bound of the mutual information between the learned representation and the data labels. Moreover, considering the $h_{i,\theta}$, the mutual information between the neighbor set and the label of anchor node v_i has a lower bound that is guaranteed by the mutual information between the anchor node and labels. The gap is determined by the homophily rate $h_{i,\theta}$ of the learned graph structure, or in other words, the learned probability $P(y_{\mathcal{N}(i)} = y_i|v_i;\theta)$. As the $h_{i,\theta}$ increases, the lower bound of $I(f(v_i);Y)$ also increases.

In summary, for our NeCo method, the homophily discrimination module for learning and updating the neighbor set $N(v_i;\theta)$ plays a crucial role in improving the performance of GCL and mining neighbor information. It also implies the necessity of the regularization term in Equation 5, because the homophily rate h will not be improved without discarding the heterophily edges.

Moreover, it has been proposed that the essence of GCL is instance discrimination as a pretext task for self-supervised learning (Wang et al., 2022; Liu et al., 2020). However, learning semantics relies heavily on intra-class data distribution, as shown in (Wang et al., 2022). It is limited by the one-to-one positive sampling strategy, and the NeCo may improve it to form a new task that we referred to as intra-class **group discrimination task**, which has less reliance on the data distribution. Here we explain why NeCo is equivalent to the group discrimination task which surpasses the existing instance discrimination tasks.

Theorem 3. *Suppose a graph is divided into k groups according to the learned structure connection. Intra-group nodes will be encoded to similar embeddings, and the training objective is equivalent to k-group discrimination as given:*

$$-l \iff \sum_{k\in c(i)} sim(z_i, z_k) - \sum_{j\notin c(i)} sim(z_i, z_j) \quad (8)$$

It can be observed that the groups c generated by $P(y_{\mathcal{N}(i)} = y_i \mid v_i\,;\,\theta)$ and the graph structure can be distinguishable. Compared to the common instance discrimination task that GCL usually adopts, group discrimination takes advantages to node representation learning. First, the label consistency assumption is no longer required, and GCL methods do not have to bear the possible semantic damage introduced by GDAs. Second, we consider the impact of the self-supervised GCL training on downstream fine-tuning. The instance discrimination task learns $|\mathcal{V}|$ node embeddings z, each of which is uniformly distributed, and the downstream classifier maps them to $|\mathcal{Y}|$ labels. But the group discrimination task learns k groups to map to labels. It can be inferred that the entropy $H(z_i|y_i)$ is larger than $H(z_c|y_c)$ learned by group discrimination since $|\mathcal{Y}| \leq k \leq |\mathcal{N}|$, and the downstream networks may encounter greater difficulties in mapping them to the corresponding labels in traditional GCLs.

Suppose that θ can distinguish all heterophily edges and that nodes with the same label in the graph are all connected, NeCo can classify nodes without a downstream task network since groups $\{c\}$ must be injective on $\mathcal{Y}$. At the same time, GNNs will achieve the same classification result as GCL with message passing on the learned topology of NeCo. Therefore, our neighboring sampling contrastive design is superior to the existing GCL methods.

4. Experiments

In this section, we conduct a wide range of experiments to demonstrate the superiority of our neighboring sampling strategy and the proposed NeCo framework. First, we evaluate the performance under the task of node classification and analyze the effect of hyperparameters. We then observe how the homophily structure that NeCo optimized facilitates GNNs.

4.1. Results of Node Classification

Settings and Datasets. GCN (Kipf & Welling, 2017) is the base encoder to extract node representations for the baselines and our proposed model. Models are trained in a fully unsupervised way, and the learned embeddings are used to train a simple L2-regularized logistic regression classifier. The homophily datasets we used are citation networks, including Cora, Citeseer, Pubmed and DBLP, and the heterophily datasets consist of web page networks (Cornell, Texas and Wisconsin), Actor and Wikipedia network Chameleon. For the homophily networks, we adopted the commonly-used 10%/10%/80% nodes for training, validation and testing. We changed the data split for heterophily datasets to 60%/20%/20%, which follows the existing GNN works on heterophily problems since strong homophily datasets contain rich label information so that during message passing. In contrast, the heterophily dataset needs more labels to fine-tune the downstream classifier. A summary of datasets with details is provided in the Appendix. Note that for the baseline AD-GCL, instead of directly using the dual-view model design in (Suresh et al., 2021), we migrate its min-max idea to our neighbor contrastive framework and name it as AD-NeCo, which updates the neighbor sets by maximizing the loss in Equation 5.

Observation. Table 1 shows the results for node classification on homophily networks. Although the homophily rate h of these datasets is relatively high, meaning that the message passing and positive samplings in the dual-branch baselines (GRACE, GCA and BGRL) will not be significantly affected, our proposed NeCo still outperforms the baselines in most cases. The dynamic training process of NeCo is shown in Figure 3(c). The homophily discrimination module tends to sample all edges at first to reduce loss. As the semantic learning of GCL progresses, NeCo identifies edges that are not useful to GNNs, and the loss objective will decrease as more edges are dropped.

For the results on heterophily networks, as shown in Table 2, we observed a vast performance degeneration of GCL methods on these tasks. Compared to the homophily datasets, the performance gap between the untrained GCN and the GCL methods is considerably narrowed. We make the following analysis. First, as shown in (Wang et al., 2022), the intra-class semantic alignment of embeddings learned by the one-to-one sampling GCL model largely depends on the data distribution. However, data's topology and attribute distribution cannot be guaranteed for heterophily networks so traditional frameworks will fail and even be worse than the randomly initialized encoder. Second, the heterophily structure of the graph will harm the message passing of encoders in the GCL model, which leads to performance degradation. Third, all baselines adopt GDAs to create views. But according to our analysis above, the task-relevant information in the heterophily graph is less than that in the strong homophily graphs. Hence GDAs are more likely to damage the semantic information and lead to lower performance.

For AD-NeCo, the proposed min-max principle is to operate aggressive samplings to make the larger difference between contrasting pairs as possible. Compared to the graph-level task in (Suresh et al., 2021), this idea does not perform well on node-level classification tasks. We make the following analysis. First, min-max principle and our NeCo affect the network's structure. While for homophily graphs, most of the edges and attributes are task-related, so the aggressive learning method brings large semantic distortion and leads to insufficient encoding. Second, the homophily discrimination of NeCo aims to keep edges with node pairs that have the same semantic labels, which is not in the opposite direction of the optimization goal of GCL as AD-GCL does. Using a single minimization objective to update the parameters is more reasonable than the min-max principle. Furthermore, unlike ADGCL, our NeCo uses the results of parametric samplings to update the graph's structure. Compared to individually making the structures parameterized, it

Dataset	Cora	Citeseer	Pubmed	DBLP
untrained GCN	66.34±2.65	60.78±1.60	84.37±00.64	70.63±0.48
DGI	82.60±0.40	68.80±0.70	86.00±0.10	83.20±0.10
GRACE	83.30±0.40	72.10±0.50	**86.70±0.10**	84.20±0.10
GCA	82.90±0.41	72.14±0.06	86.01±0.05	84.06±0.02
BGRL	82.77±0.75	68.45±0.15	84.34±0.17	80.63±0.46
AD-NeCo	79.56±0.19	68.71±0.38	82.05±0.65	80.73±0.21
NeCo	**83.84±0.54**	**73.06±0.93**	86.29±0.21	**84.45±0.19**
supervised GCN	82.80	72.00	84.90	82.70

Table 1. Node classification results on homophily datasets.

	Cornell	Texas	Wisconsin	Actor	chameleon
untrained GCN	49.55±7.75	49.55±5.10	46.41±1.85	28.93±1.36	42.03±0.92
DGI	52.25±7.09	54.56±6.74	54.90±2.77	27.87±0.89	42.91±2.47
GRACE	53.15±7.75	55.86±3.37	49.02±6.98	29.78±0.51	40.94±3.67
BGRL	51.25±0.85	54.20±1.60	48.44±1.43	25.35±0.27	40.07±1.37
AD-NeCo	56.14±4.95	55.83±1.14	53.72±2.96	29.97±0.93	41.42±1.61
NeCo	**59.36±4.59**	**59.45±2.55**	**57.20±4.23**	**30.78±1.02**	**44.23±1.13**
supervised GCN	55.14	55.68	53.73	30.64	28.18

Table 2. Node classification results on heterophily datasets.

is more appropriate for node-level tasks to strengthen structures with the help of the semantic learning capabilities of GCL methods.

4.2. Solving the GNN Problem

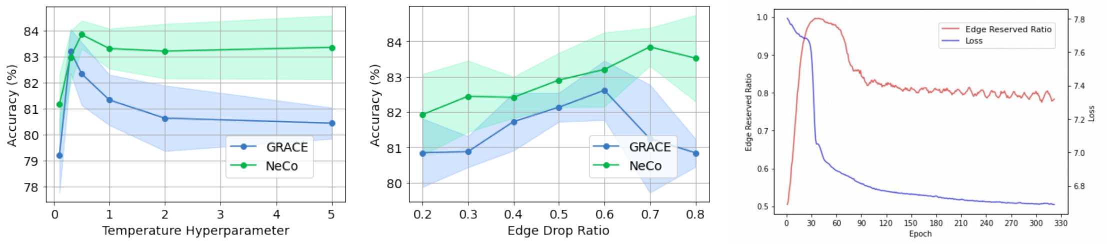
Here we experimentally demonstrate that the learned structure benefits the learning of GNNs. We select the Sigmoid(·) instead of the Gumbel-Max trick following the homophily discrimination module to get the existence probability of edges. Then, we set the threshold to get the learned structure. We chose GraphSAGE (Hamilton et al., 2017) (a classic method that achieves strong performance on heterophily datasets) as the baseline. We compared the result trained with the input graph and the structure NeCo learned. We believe that the neighbor sampling of GraphSAGE is the key to solving the homophily problem, and the graph structure we learned improves the homophily rate h of the graph. Therefore, GraphSAGE can sample more nodes with the same label when $P(y_{\mathcal{N}(i)} = y_i|v_i;\theta)$ is stronger, leading to performance increase.

4.3. Analysis of Hyperparameters

Here we study how the different temperature hyperparameter τ and edge drop ratio p affect the training of our proposed NeCo. We vary the hyperparameter τ within the scope of {0.1,0.3,0.5,1,2,5} to observe the accuracy variation of NeCo and GRACE on the test set. For the drop ratio p, we modify the regularization term in Equation 5 to $\lambda(||\sum_{i\in\mathcal{V}}\mathcal{N}(i;\theta)||/|\mathcal{E}|-p)$ to control the proportion of edges that NeCo drops. As shown in Figure 3, GRACE shows significant fluctuation when facing the changing temperature τ (about 5% on Cora). On the contrary, NeCo offers good stability to the change of temperature parameters and outperforms GRACE in all settings. This indicates that the neighbor-based sampling strategy is less sensitive to the τ than the one-to-one sampling strategy. For the edge drop ratio p, GRACE will be problematic with message passing when edges are dropped a lot, causing the intra-class data distribution to be corrupted and the performance degradation. But NeCo could keep learning good semantics of nodes through sampling between the selected neighbor sets because of the dynamic edge adjustment process, as shown in Figure 3.

5. Related Work

Self-supervised learning has become a powerful method for representation learning on unlabeled data, and contrastive learning is the most representative and successful paradigm. It was first proposed in the representation learning of computer vision, following the mutual information maximization principle (Belghazi et al., 2018) and achieving competitive performance to the supervised models. Many inspirations of them, such as mutual information estimation (Hjelm et al., 2019), dense InfoNCE contrasting (Chen et al., 2020), asymmetric design (Grill et al., 2020) and feature de-correlation (Zbontar et al., 2021), have also been verified to be applicable with a strong performance in graph learning on node level tasks (Zhu et al., 2020b; Thakoor et al., 2021; Bielak et al., 2021; Velickovic et al., 2019) and



(a) (b) (c)

Figure 3. (a)(b) The effect of temperature parameter τ and drop ratio p on the one-to-one sampling model GRACE and the neighbor sampling model NeCo. (c) The dynamic training process of NeCo.

Datasets	Cornell	Texas	Wisconsin	Actor	chameleon
GraphSAGE	70.84±5.27	80.41±5.16	78.49±3.53	34.08±1.52	42.45±1.97
NeCo+GraphSAGE	75.38±4.02	82.29±4.14	82.72±2.10	34.99±0.79	45.16±2.21

Table 3. Semi-supervised node classification results of GraphSAGE and the model augmented with NeCo.

graph level tasks (Sun et al., 2020; You et al., 2020; Zhang et al., 2020). Most existing GCL methods that build on these ideas aims at exploring new view designs dominated by prior domain knowledge or implemented with additional parameters. MVGRL (Hassani & Ahmadi, 2020) and GCA (Zhu et al., 2021c) use graph diffusion and node centrality as the principle for performing GDAs. JOAO (You et al., 2021) and AD-GCL (Suresh et al., 2021) use additional parametric components to control the type and distribution of GDAs for better downstream performance. (Xia et al., 2022; Yu et al., 2021a) add noises based on the prior distribution to the embeddings or the parameters of the encoder to create multiple views, which is an implicit way to perform GDAs. Note that they all follow the fixed paradigm of visual contrastive learning, where the one-to-one samplings are imprisoned in the same index between augmented views.

In contrast, our proposed NeCo with sufficient theoretical proof optimizes the positive samplings strategy for GCL. While many state-of-the-art GCL models (Kefato et al., 2021a; Suresh et al., 2021) use parameters to control data augmentation, there are many fundamental differences: 1) They still follow a rigid inter-view one-to-one contrasting with the parametric GDAs or views. However, NeCo does not require additional views or GDAs, and the parametric neighbor sampling could benefit the GCL more than the traditional strategy. 2) AD-GCL (Suresh et al., 2021) is based on the information bottleneck theory, and it maximizes the loss between the original view and GDAs to learn the augmenter. Our proposed NeCo is not a min-max issue, and it learns the homophily discriminator from the positive samplings of GCL instead of updating parameters by augmenting graphs like AD-GCL.

Some efforts have also been dedicated to solving the problem of generalizing GNNs to graphs with low homophily in specific downstream tasks (Jin et al., 2023). They can be divided into two mainstream types of methods depending on the aggregation mechanism: (1) (Yan et al., 2021; Zhu et al., 2021a; Yang et al., 2021; Wang et al., 2021; He et al., 2021) all adopt the traditional first-order neighbor aggregation mechanism but use heterophily information to learn weights for neighbors with different labels, which is similar to our proposed NeCo solution. (2) (Pei et al., 2020; Zhu et al., 2020a; Chien et al., 2021; Lim et al., 2021) believe that for weak homophily networks, intra-class nodes are more likely to appear in higher-order neighbors, where more useful information could be propagated in GNNs compared to the traditional first-order aggregation mechanism. They all follow supervised learning styles, and to the best of our knowledge, NeCo is the first to explore homophily problems with changing structure in a self-supervised way

6. Conclusion

In this paper, we integrated the positive sampling strategy of GCL and the homophily discrimination of GNNs in the same framework and developed a new idea of solving them simultaneously. We have proposed to extend the range of positive samplings to node neighbor sets. It allows us to develop a new paradigm of contrastive learning to avoid creating extra views in the traditional contrastive models and to eliminate data augmentation completely. To address the inter-class node pairs in neighbor sets, we proposed a parametric homophily discrimination module. It learns and updates the intra-class neighbor sets by affecting the objective of neighbor sampling for GCL and graph encoding of GNNs, that is, to improve the homophily rate of the structure. We theoretically prove that NeCo can guarantee a lower bound of mutual information and the homophily rate h increased with training improves this lower bound. Combined with the learned neighbor sets, our sampling strategy can ensure that the representation of $N(v_i;\theta)$ captures certain information with the downstream tasks. Moreover, the group discrimination task that NeCo achieves facilities fine-tuning for downstream tasks. We believe this is the first to propose an unsupervised method for solving the homophily problem with the help of the solid semantic learning capability introduced by GCL.

7. Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. U22B2036, 61832014, 62276187, 62272340, 11931015), National Science Fund for Distinguished Young Scholars (Grants No. 62025602), Tencent Foundation, XPLORER PRIZE, and the Hong Kong Global STEM Professorship scheme and the Hong Kong Jockey Club Charities Trust.

References

Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. C. Mine: Mutual information neural estimation. *ArXiv*, abs/1801.04062, 2018.

Bielak, P., Kajdanowicz, T., and Chawla, N. Graph barlow twins: A self-supervised representation learning framework for graphs. *ArXiv*, abs/2106.02466, 2021.

Cai, H., Zheng, V. W., and Chang, K. C. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, 30(9):1616–1637, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

Chien, E., Peng, J., Li, P., and Milenkovic, O. Adaptive universal generalized pagerank graph neural network. *arXiv: Learning*, 2021.

Duvenaud, D. K., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T. D., Aspuru-Guzik, A., and Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *ArXiv*, abs/1509.09292, 2015.

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., and Yin, D. Graph neural networks for social recommendation. In *WWW*, pp. 417–426, 2019.

Grill, J.-B., Strub, F., Altch'e, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv*, abs/2006.07733, 2020.

Hamilton, W. L., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *NIPS*, 2017.

Hassani, K. and Ahmadi, A. H. K. Contrastive multi-view representation learning on graphs. *ArXiv*, abs/2006.05582, 2020.

He, D., Liang, C., Liu, H., Wen, M.-C., Jiao, P., and Feng, Z. Block modeling-guided graph convolutional neural networks. *ArXiv*, abs/2112.13507, 2021.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *ArXiv*, abs/1808.06670, 2019.

Jin, D., Huo, C., Liang, C., and Yang, L. Heterogeneous graph neural network via attribute completion. In *Proceedings of the Web Conference 2021*, pp. 391–400, 2021a.

Jin, D., Yu, Z., Jiao, P., Pan, S., He, D., Wu, J., Yu, P. S., and Zhang, W. A survey of community detection approaches: From statistical modeling to deep learning. *IEEE Trans. Knowl. Data Eng.*, 35(2):1149–1170, 2023. doi: 10.1109/TKDE.2021.3104155. URL https://doi.org/10.1109/TKDE.2021.3104155.

Jin, M., Zheng, Y., Li, Y.-F., Gong, C., Zhou, C., and Pan, S. Multi-scale contrastive siamese networks for self-supervised graph representation learning. In *IJCAI*, 2021b.

Jing, L. and Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:4037–4058, 2021.

Kefato, Z. T., Girdzijauskas, S., and Stärk, H. Jointly learnable data augmentations for self-supervised gnns. *ArXiv*, abs/2108.10420, 2021a.

Kefato, Z. T., Girdzijauskas, S., and Stärk, H. Jointly learnable data augmentations for self-supervised gnns. *ArXiv*, abs/2108.10420, 2021b.

Kipf, T. and Welling, M. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907, 2017.

Lim, D., Hohne, F., Li, X., Huang, S., Gupta, V., Bhalerao, O., and Lim, S.-N. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *ArXiv*, abs/2110.14446, 2021.

Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *ArXiv*, abs/2006.08218, 2020.

Liu, Y., Pan, S., Jin, M., Zhou, C., Xia, F., and Yu, P. S. Graph self-supervised learning: A survey. *ArXiv*, abs/2103.00111, 2022.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *ArXiv*, abs/1611.00712, 2017.

Pei, H., Wei, B., Chang, K. C.-C., Lei, Y., and Yang, B. Geom-gcn: Geometric graph convolutional networks. *ArXiv*, abs/2002.05287, 2020.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009.

Schmarje, L., Santarossa, M., Schröder, S.-M., and Koch, R. A survey on semi-, self- and unsupervised techniques in image classification. *ArXiv*, abs/2002.08721, 2020.

Sun, F.-Y., Hoffmann, J., and Tang, J. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *ArXiv*, abs/1908.01000, 2020.

Suresh, S., Li, P., Hao, C., and Neville, J. Adversarial graph augmentation to improve graph contrastive learning. *ArXiv*, abs/2106.05819, 2021.

Thakoor, S., Tallec, C., Azar, M. G., Munos, R., Velivcković'c, P., and Valko, M. Bootstrapped representation learning on graphs. *ArXiv*, abs/2102.06514, 2021.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio', P., and Bengio, Y. Graph attention networks. *ArXiv*, abs/1710.10903, 2018.

Velickovic, P., Fedus, W., Hamilton, W. L., Lio', P., Bengio, Y., and Hjelm, R. D. Deep graph infomax. *ArXiv*, abs/1809.10341, 2019.

Wang, T., Wang, R., Jin, D., He, D., and Huang, Y. Powerful graph convolutioal networks with adaptive propagation mechanism for homophily and heterophily. *ArXiv*, abs/2112.13562, 2021.

Wang, Y., Zhang, Q., Wang, Y., Yang, J., and Lin, Z. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *ArXiv*, abs/2203.13457, 2022.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. *ArXiv*, abs/2010.03622, 2021.

Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., and Tan, T. Session-based recommendation with graph neural networks. In *AAAI*, pp. 346–353, 2019.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(1):4–24, 2021.

Xia, J., Wu, L., Chen, J., Hu, B., and Li, S. Z. Simgrace: A simple framework for graph contrastive learning without data augmentation. *Proceedings of the ACM Web Conference 2022*, 2022.

Xu, D., Cheng, W., Luo, D., Chen, H., and Zhang, X. Infogcl: Information-aware graph contrastive learning. *ArXiv*, abs/2110.15438, 2021.

Yan, Y., Hashemi, M., Swersky, K., Yang, Y., and Koutra, D. Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks. *ArXiv*, abs/2102.06462, 2021.

Yang, L., Li, M., Liu, L., Niu, B., Wang, C., Cao, X., and Guo, Y. Diverse message passing for attribute with heterophily. 2021.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph contrastive learning with augmentations. *ArXiv*, abs/2010.13902, 2020.

You, Y., Chen, T., Shen, Y., and Wang, Z. Graph contrastive learning automated. In *ICML*, 2021.

Yu, J., Yin, H., Xia, X., zhen Cui, L., and Nguyen, Q. V. H. Graph augmentation-free contrastive learning for recommendation. *ArXiv*, abs/2112.08679, 2021a.

Yu, Z., Jin, D., Liu, Z., He, D., Wang, X., Tong, H., and Han, J. AS-GCN: adaptive semantic architecture of graph convolutional networks for text-rich networks. In Bailey, J., Miettinen, P., Koh, Y. S., Tao, D., and Wu, X. (eds.), *IEEE International Conference on Data Mining, ICDM 2021, Auckland, New Zealand, December 7-10, 2021*, pp. 837–846. IEEE, 2021b. doi: 10.1109/ICDM51629.2021.00095. URL https://doi.org/10.1109/ICDM51629.2021.00095.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Zhang, H., Lin, S., Liu, W., Zhou, P., Tang, J., Liang, X., and Xing, E. P. Iterative graph self-distillation. *ArXiv*, abs/2010.12609, 2020.

Zheng, X., Liu, Y., Pan, S., Zhang, M., Jin, D., and Yu, P. S. Graph neural networks for graphs with heterophily: A survey. *arXiv preprint arXiv:2202.07082*, 2022.

Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., and Koutra, D. Beyond homophily in graph neural networks: Current limitations and effective designs. *arXiv: Learning*, 2020a.

Zhu, J., Rossi, R. A., Rao, A. B., Mai, T., Lipka, N., Ahmed, N., and Koutra, D. Graph neural networks with heterophily. In *AAAI*, 2021a.

Zhu, W., Wang, X., and Cui, P. Deep learning for learning graph representations. *ArXiv*, abs/2001.00293, 2019.

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., and Wang, L. Deep graph contrastive representation learning. *ArXiv*, abs/2006.04131, 2020b.

Zhu, Y., Xu, Y., Liu, Q., and Wu, S. An empirical study of graph contrastive learning. *ArXiv*, abs/2109.01116, 2021b.

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., and Wang, L. Graph contrastive learning with adaptive augmentation. *Proceedings of the Web Conference 2021*, 2021c.

A. Limitations and Broader Impact

We focused on two fundamentally important problems of graph deep learning, the extreme imbalance between graph data and labels and the extensive heterophily noise problem that affects message passing in graph neural networks. The proposed method can reduce the influence of heterophily information under the premise of unsupervised learning. It also accommodates well the downstream self-supervised learning tasks. The experiments and theoretical analysis carried out that this research are expected to inspire future development in graph/node learning.

B. Theoretical Proofs

B.1. Proof of Theorem 1

Theorem 1. *Suppose that each node v_i has a homophily rate h_i, which could be defined as $h_i = |\mathcal{N}(i) \cap \mathcal{Y}(i)|/degree_i$. Our method maximizes the lower bound of the mutual information between the representations of the contrastive pairs in NeCo, and the gap is determined by the learned homophily rate $h_{i,\theta}$ of the graph G. Specifically,*

$$-L \leq I(f(v_i), f(\mathcal{N}(i;\theta))) + \mathbb{E}_{P(v_i)}(h_{i;\theta} - \frac{1}{|\mathcal{N}(i;\theta)|}) \sum sim(z_i, z_{\mathcal{N}(i;\theta)})/\tau \tag{9}$$

Proof. First, we review the loss of our proposed NeCo method and rewrite the loss for every node v_i as:

$$-l_i = \log \sum_{k \in \mathcal{N}(i;\theta)} \exp\left(sim(z_i, z_k)/\tau\right) - \log \sum_{j \in \{\mathcal{V} - \mathcal{N}(i;\theta)\}} \exp\left(sim(z_i, z_j)/\tau\right) \tag{10}$$

where $\mathcal{N}(i;\theta)$ indicates the sampling result by the Gumbel-Max trick with the probability $P(y_{\mathcal{N}(i)} = y_i | v_i; \theta)$. We further define the $\mathcal{N}_+(i) = \{\mathcal{N}(i) \cap Y(i)\}$ to indicate the neighbor sets with the same label and the complementary set as $\mathcal{N}_-(i) = \{\mathcal{N}(i) - \mathcal{N}_+(i)\}$. For the first term in Equation 10, we can get:

$$\begin{aligned}
item &= \log \sum_{k \in \mathcal{N}(i;\theta)} \exp\left(sim(z_i, z_k)/\tau\right) \\
&= \log\Big(\sum_{k \in \mathcal{N}_+(i;\theta)} \exp\left(sim(z_i, z_k)/\tau\right) + \sum_{m \in \mathcal{N}_-(i;\theta)} \exp\left(sim(z_i, z_m)/\tau\right)\Big) \\
&\leq \log\Big(\sum_{k \in \mathcal{N}(i;\theta)} \frac{\mathbb{1}_{k \in \mathcal{N}_+(i;\theta)}}{|\mathcal{N}(i;\theta)|} \exp\left(sim(z_i, z_k)/\tau\right)\Big) + \log|\mathcal{N}(i;\theta)| \\
&\leq h_{i;\theta} \sum_{k \in \mathcal{N}(i;\theta)} \log\big(\exp\left(sim(z_i, z_k)/\tau\right)\big) \\
&= h_{i;\theta} \sum_{k \in \mathcal{N}(i;\theta)} (sim(z_i, z_k)/\tau)
\end{aligned} \tag{11}$$

where the second inequality holds because of Jensen's inequality and the expectation $\sum \frac{\mathbb{1}_{k \in \mathcal{N}_+(i)}}{|\mathcal{N}(i)|}$ is exactly the same as the definition of homophily rate h_i. Then we can get the loss in the form:

$$-l_i \leq h_{i;\theta} \sum_{k \in \mathcal{N}(i;\theta)} (sim(z_i, z_k)/\tau) - \log \sum_{j \in \{\mathcal{V} - \mathcal{N}(i;\theta)\}} \exp\left(sim(z_i, z_j)/\tau\right) \tag{12}$$

Then the loss for the entire graph can be expressed as:

$$-L \leq \frac{1}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} h_{i;\theta} \sum_{k \in \mathcal{N}(i;\theta)} (sim(z_i, z_k)/\tau) - \frac{1}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} \log \sum_{j \in \{\mathcal{V} - \mathcal{N}(i;\theta)\}} \exp\left(sim(z_i, z_j)/\tau\right) \tag{13}$$

We rewrite it in the expectation form and replace the $sim(z_i, z_j)/\tau$ to $g(v_i, v_j)$ as:

$$
\begin{aligned}
-L &\leq \mathbb{E}_{P(v_i)}(|\mathcal{N}(i;\theta)|h_{i;\theta})\mathbb{E}_{P(\mathcal{N}(i;\theta)|v_i)}g(v_i, v_+) - \mathbb{E}_{P(v_i)}\log(\mathbb{E}_{P(v_-)}\exp(g(v_i, v_-)) \\
&= \mathbb{E}_{P(v_i)}(h_{i;\theta} - \frac{1}{|\mathcal{N}(i;\theta)|})\sum_{k\in\mathcal{N}(i;\theta)} g(v_i, v_+) + \mathbb{E}_{P(v_i,\mathcal{N}(i;\theta))}g(v_i, v_+) - \mathbb{E}_{P(v_i)}\log(\mathbb{E}_{P(v_-)}\exp(g(v_i, v_-)) \\
&= I(v_i, \mathcal{N}(i;\theta)) + \mathbb{E}_{P(v_i)}(h_{i;\theta} - \frac{1}{|\mathcal{N}(i;\theta)|})\sum_{k\in\mathcal{N}(i;\theta)} g(v_i, v_+)
\end{aligned}
\tag{14}
$$

which is the lower bound.

B.2. Proof of Theorem 2

Theorem 2. *With the defined homophily rate h_i for each node, the mutual information between the representations and the downstream task labels Y can be written as:*

$$
I(f(v_i); Y) \geq I(N(v_i;\theta); Y) = I(v_i; Y) + (1 - h_{i;\theta})\big(H(Y|v_i) - \sum_{y\in Y}\frac{\log(|\mathcal{Y}| - 1)}{|\mathcal{Y}| - 1}\big)
\tag{15}
$$

Proof. According to the conclusion in AD-GCL and the data processing inequality, we have:

$$
I(f(v_i); Y) \geq I(f(N(v_i;\theta)); Y) = I(N(v_i;\theta); Y)
\tag{16}
$$

because the parametric sampling $N(v_i;\theta)$ in NeCo can be regarded as the natural optimal data augmentation based on the data itself. Then based on the definition of mutual information we have:

$$
I(N(v_i;\theta); Y) - I(v_i; Y) = H(Y|v_i) - H(Y|N(v_i;\theta))
\tag{17}
$$

We next expand the second term with the definition of information entropy and get:

$$
\begin{aligned}
H(Y|N(v_i;\theta)) &= -\sum_{y\in Y} P(y|N(v_i;\theta))\log P(y|N(v_i;\theta)) \\
&= -h_{i;\theta}\sum_{y\in Y} P(y|v_i))\log P(y|v_i) - (1 - h_{i;\theta})\sum_{y\in Y} P(y|N_-(v_i;\theta))\log P(y|N_-(v_i;\theta)) \\
&= h_{i;\theta}H(Y|v_i) - (1 - h_{i;\theta})\sum_{y\in Y}\frac{1}{|\mathcal{Y}| - 1}\log\frac{1}{|\mathcal{Y}| - 1} \\
&= h_{i;\theta}H(Y|v_i) + (1 - h_{i;\theta})\sum_{y\in Y}\frac{\log(|\mathcal{Y}| - 1)}{|\mathcal{Y}| - 1}
\end{aligned}
\tag{18}
$$

Here we treat the homophily rate $h_{i;\theta}$ as a probability rather than an observed statistics in the Bayesian analysis. The second equation holds because for a node v_i, the probability of its neighbor having the same label with node v_i is $h_{i;\theta}$. For the remaining $|\mathcal{Y}| - 1$ labels, we assume that it follows a uniform distribution and the probability for each label is $\frac{1}{|\mathcal{Y}|-1}$.

Combining the above formulas, we have:

$$
\begin{aligned}
I(N(v_i;\theta); Y) &= I(v_i; Y) + H(Y|v_i) - H(Y|N(v_i;\theta)) \\
&= I(v_i; Y) + (1 - h_{i;\theta})H(Y|v_i) - (1 - h_{i;\theta})\sum_{y\in Y}\frac{\log|\mathcal{Y}| - 1}{|\mathcal{Y}| - 1} \\
&= I(v_i; Y) + (1 - h_{i;\theta})\big(H(Y|v_i) - \sum_{y\in Y}\frac{\log(|\mathcal{Y}| - 1)}{|\mathcal{Y}| - 1}\big)
\end{aligned}
\tag{19}
$$

Q.E.D.

B.3. Proof of Theorem 3

Theorem 3. *Suppose that a graph is divided into k groups according to the learned structure connection. Intra-group nodes will be encoded to similar embeddings, and the training objective is equivalent to a k-group discrimination as given:*

$$-l \iff \sum_{k\in c(i)} sim(z_i, z_k) - \sum_{j\notin c(i)} sim(z_i, z_j) \tag{20}$$

Proof. We rewrite the loss function of node v_i with Taylor expansion of the first order as:

$$\begin{aligned} l_i &= \log \frac{\sum_{j\in\{\mathcal{V}-\mathcal{N}(i;\theta)\}} \exp\left(sim(z_i, z_j)/\tau\right)}{\sum_{k\in\mathcal{N}(i;\theta)} \exp\left(sim(z_i, z_k)/\tau\right)} \\ &= \log(1 + \frac{\sum_{j\in\{\mathcal{V}-\mathcal{N}(i;\theta)\}} \exp\left(sim(z_i, z_j)/\tau\right) - \sum_{k\in\mathcal{N}(i;\theta)} \exp\left(sim(z_i, z_k)/\tau\right)}{\sum_{k\in\mathcal{N}(i;\theta)} \exp\left(sim(z_i, z_k)/\tau\right)}) \\ &\approx \frac{\sum_{j\in\{\mathcal{V}\}} \exp\left(sim(z_i, z_j)/\tau\right)}{\sum_{k\in\mathcal{N}(i;\theta)} \exp\left(sim(z_i, z_k)/\tau\right)} - 2 \\ &\propto \sum_{j\in\{\mathcal{V}-\mathcal{N}(i;\theta)\}} \exp\left(sim(z_i, z_j)/\tau\right) - \sum_{k\in\mathcal{N}(i;\theta)} \exp\left(sim(z_i, z_k)/\tau\right) \end{aligned} \tag{21}$$

We extend the loss function of v_i to the form of the entire graph G and remove the temperature hyperparamter τ:

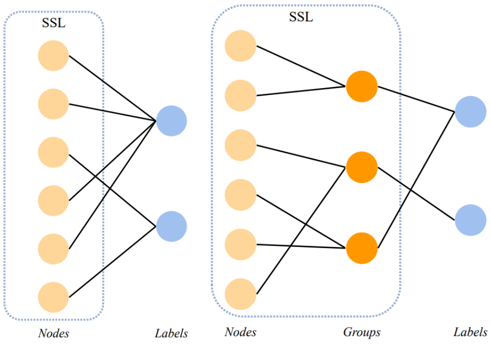
$$\begin{aligned} -L &= -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} l_i \\ &\propto \sum_{i\in\mathcal{V}} \sum_{k\in\mathcal{N}(i;\theta)} \exp\left(sim(z_i, z_k)\right) - \sum_{i\in\mathcal{V}} \sum_{j\in\{\mathcal{V}-\mathcal{N}(i;\theta)\}} \exp\left(sim(z_i, z_j)\right) \end{aligned} \tag{22}$$

The connection of graphs is defined as that there exists paths between node pairs. Nodes in G are connected by edges and G can be divided into groups based on connectivity. We can integrate the summation of all nodes v_i with the summation of the learned neighbors $\mathcal{N}(i;\theta)$ in the first item. We make the conclusion that the connected nodes form a group c, and the loss objective is to make the node embeddings within the same group tend to be consistent.

Then the graph contrastive learning method is no longer an **instance discrimination task** where nodes from different views are learned to be encoded with the consistent representation and the representation of each node are learned to be distinguishable from other nodes. NeCo improves it with respect to the **group discrimination task**. It divides nodes into k groups according to the learned topology and keeps learning consistent representations among nodes within a group. Similarly, representations that from different groups need to be distinct. Then we can rewrite the loss objective for every node v_i as:

$$-l_i \iff \sum_{k\in c(i)} sim(z_i, z_k) - \sum_{j\notin c(i)} sim(z_i, z_j) \tag{23}$$

The promotion from node-level tasks to group-level tasks is the key to making downstream classifier more efficient as shown in Figure 4.



(a) (b)

Figure 4. (a) Fine-tuning on the instance discrimination task. (b) Fine-tuning on the proposed group discrimination task.

C. Experimental Setup

All our experiments were performed on the Google Colab platform with a Tesla NVIDIA Tesla P100 (16GB) GPU. All datasets are available in *PyTorch Geometric* library.

D. Discussion on Information Bottleneck and Homophily

The information bottleneck theory and the homophily problem in GNNs are two important and related aspects of graph representation learning. In this research, we aimed to solve the problem of homophily while guaranteeing the information bottleneck constraints. The objective of graph information bottleneck can be written as:

$$\underset{f}{max}\, I(f(G);Y) - \alpha I(G;f(G)) \tag{24}$$

The core idea of our method is to capture the minimal sufficient graph information in GNNs for the downstream tasks. Specifically, we conjectured that the training process of GNNs has two separate phases: 1) an initial fitting phase that increases $I(f(G);Y)$, and 2) a subsequent compression phase that decreases $I(G;f(G))$.

For the message passing of GNNs, the homophily information is defined by the label information of nodes. First, because of the aggregation mechanism of GNNs, propagation between intra-class nodes makes the learned node distribution closer to the robust distribution of labels and the redundant label-independent distributions from intra-class nodes will be smoothed. Therefore, the information bottleneck can be measured by the label homophily rate h. Second, the dimensions of attribute vectors which are consistent in values with their neighbors would not be affected in message passing, for example, GCN adds and averages features. The dimensions with different values in the vector form can be regarded as the label-independent noise information, which can be measured by the attribute homophily rate β. It can be concluded that the homophily problem rests on the graph information bottleneck theory, and defines graph information that contributes to learning label-relevant distribution.

Therefore, the homophily rate h and β defined in Section 2.2 actually represent a measure of the information bottleneck theory. The improvement of h and β introduced in our proposed NeCo method can also meet the requirement of graph information bottleneck theory.