

# Node Similarity Preserving Graph Convolutional Networks

Wei Jin  
Michigan State University  
jinwei2@msu.edu

Tyler Derr  
Vanderbilt University  
tyler.derr@vanderbilt.edu

Yiqi Wang  
Michigan State University  
wangy206@msu.edu

Yao Ma  
Michigan State University  
mayao4@msu.edu

Zitao Liu  
TAL Education Group  
liuzitao@100tal.com

Jiliang Tang  
Michigan State University  
tangjili@msu.edu

## ABSTRACT

Graph Neural Networks (GNNs) have achieved tremendous success in various real-world applications due to their strong ability in graph representation learning. GNNs explore the graph structure and node features by aggregating and transforming information within node neighborhoods. However, through theoretical and empirical analysis, we reveal that the aggregation process of GNNs tends to destroy node similarity in the original feature space. There are many scenarios where node similarity plays a crucial role. Thus, it has motivated the proposed framework SimP-GCN that can effectively and efficiently preserve node similarity while exploiting graph structure. Specifically, to balance information from graph structure and node features, we propose a feature similarity preserving aggregation which adaptively integrates graph structure and node features. Furthermore, we employ self-supervised learning to explicitly capture the complex feature similarity and dissimilarity relations between nodes. We validate the effectiveness of SimP-GCN on seven benchmark datasets including three assortative and four disassortative graphs. The results demonstrate that SimP-GCN outperforms representative baselines. Further probe shows various advantages of the proposed framework. The implementation of SimP-GCN is available at <https://github.com/ChandlerBang/SimP-GCN>.

## ACM Reference Format:

Wei Jin, Tyler Derr, Yiqi Wang, Yao Ma, Zitao Liu, and Jiliang Tang. 2020. Node Similarity Preserving Graph Convolutional Networks. In *Proceedings of 14th ACM International Conference On Web Search And Data Mining (WSDM 2021)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Graphs are essential data structures that describe pairwise relations between entities for real-world data from numerous domains such as social media, transportation, linguistics and chemistry [1, 24, 35, 42]. Many important tasks on graphs involve predictions over nodes and edges. For example, in node classification, we aim to predict labels of unlabeled nodes [15, 17, 33]; and in link prediction, we want to infer whether there exists an edge between a given

node pair [16, 41]. All these tasks can be tremendously facilitated if good node representations can be extracted.

Recent years have witnessed remarkable achievements in Graph Neural Networks (GNNs) for learning node representations that have advanced various tasks on graphs [1, 24, 35, 42]. GNNs explore the graph structure and node features by aggregating and transforming information within node neighborhoods. During the aggregation process, GNNs tend to smooth the node features. Note that in this work, we focus on one of the most popular GNN models, i.e., graph convolution network (or GCN) [15]; however, our study can be easily extended to other GNN models as long as they follow a neighborhood aggregation process. By connecting GCN with Laplacian smoothing, we show that the graph convolution operation in GCN is equivalent to one-step optimization of the signal recovering problem (see Section 3.1). This operation will reduce the overall node feature difference. Furthermore, through empirical study, we demonstrate that (1) given a node, nodes that have top feature similarity with the node are less likely to overlap with these that connect to the node (we refer to this as the “Less-Overlapping” problem in this work); and (2) GCN tends to preserve structural similarity rather than node similarity, no matter whether the graph is assortative (when node homophily [25] holds in the graph) or not (see Section 3.2).

Neighbor smoothing from the aggregation process and the less-overlapping problem inevitably introduce one consequence – node representations learned by GCN naturally destroy the node similarity of the original feature space. This consequence can reduce the effectiveness of the learned representations and hinder the performance of downstream tasks since node similarity can play a crucial role in many scenarios. Next we illustrate several examples of these scenarios. First, when the graph structure is not optimal, e.g., graphs are disassortative [26] or graph structures have been manipulated by adversarial attacks, relying heavily on the structure information can lead to very poor performance [13, 27, 36, 44]. Second, since low-degree nodes only have a small number of neighbors, they receive very limited information from neighborhood and GNNs often cannot perform well on those nodes [32]. This scenario is similar to the cold-start problem in recommender systems where the system cannot provide a satisfying recommendation for new users or items [2]. Third, for nodes affiliated to the “hub” nodes which presumably connect with many different types of nodes, the information from the hub nodes could be very confusing and misleading [37]. Fourth, when applying multiple graph convolution layers, GNNs can suffer over-smoothing problem that the learned node embeddings become totally indistinguishable [18, 22, 28].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

WSDM 2021, March 8–12, 2021, Jerusalem, Israel

© 2020 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

In this work, we aim to design a new graph convolution model that can better preserve the original node similarity. In essence, we are faced with two challenges. First, **how to balance the information from graph structure and node features during aggregation?** We propose an adaptive strategy that coherently integrates the graph structure and node features in a data-driven way. It enables each node to adaptively adjust the information from graph structure and node features. Second, **how to explicitly capture the complex pairwise feature similarity relations?** We employ a self-supervised learning strategy to predict the pairwise feature similarity from the hidden representations of given node pairs. By adopting similarity prediction as the self-supervised pretext task, we are allowed to explicitly encode the pairwise feature relation. Furthermore, combining the above two components leads to our proposed model, SimP-GCN, which can effectively and adaptively preserve feature and structural similarity, thus achieving state-of-the-art performance on a wide range of benchmark datasets. Our contributions can be summarized as follows:

- We theoretically and empirically show that GCN can destroy original node feature similarity.
- We propose a novel GCN model that effectively preserves feature similarity and adaptively balances the information from graph structure and node features while simultaneously leveraging their rich information.
- Extensive experiments have demonstrated that the proposed framework can outperform representative baselines on both assortative and disassortative graphs. We also show that preserving feature similarity can significantly boost the robustness of GCN against adversarial attacks.

## 2 RELATED WORK

Over the past few years, increasing efforts have been devoted toward generalizing deep learning to graph structured data in the form of graph neural networks. There are mainly two streams of graph neural networks, i.e. spectral based methods and spatial based methods [24]. Spectral based methods learn node representations based on graph spectral theory [29]. Bruna et al. [3] first generalize convolution operation to non-grid structures from spectral domain by using the graph Laplacian matrix. Following this work, ChebNet [6] utilizes Chebyshev polynomials to modulate the graph Fourier coefficients and simplify the convolution operation. The ChebNet is further simplified to GCN [15] by setting the order of the polynomial to 1 together with other approximations. While being a simplified spectral method, GCN can be also regarded as a spatial based method. From a node perspective, when updating its node representation, GCN aggregates information from its neighbors. Recently, many more spatial based methods with different designs to aggregate and transform the neighborhood information are proposed including GraphSAGE [10], MPNN [9], GAT [33], etc.

While graph neural networks have been demonstrated to be effective in many applications, their performance might be impaired when the graph structure is not optimal. For example, their performance deteriorates greatly on disassortative graphs where homophily does not hold [27] and thus the graph structure introduces noises; adversarial attack can inject carefully-crafted noise to disturb the graph structure and fool graph neural networks into

making wrong predictions [34, 44, 45]. In such cases, the graph structure information may not be optimal for graph neural networks to achieve better performance while the original features could come as a rescue if carefully utilized. Hence, it urges us to develop a new graph neural network capable of preserving node similarity in the original feature space.

## 3 PRELIMINARY STUDY

In this section, we investigate if the GCN model can preserve feature similarity via theoretical analysis and empirical study. Before that, we first introduce key notations and concepts.

Graph convolutional network (GCN) [15] was proposed to solve semi-supervised node classification problem, where only a subset of the nodes with labels. A graph is defined as  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ , where  $\mathcal{V} = \{v_1, v_2, \dots, v_n\}$  is the set of  $n$  nodes,  $\mathcal{E}$  is the set of edges describing the relations between nodes and  $\mathbf{X} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \dots, \mathbf{x}_n^\top] \in \mathbb{R}^{n \times d}$  is the node feature matrix where  $d$  is the number of features and  $\mathbf{x}_i$  indicates the node features of  $v_i$ . The graph structure can also be represented by an adjacency matrix  $\mathbf{A} \in \{0, 1\}^{n \times n}$  where  $\mathbf{A}_{ij} = 1$  indicates the existence of an edge between nodes  $v_i$  and  $v_j$ , otherwise  $\mathbf{A}_{ij} = 0$ . A single graph convolutional filter with parameter  $\theta$  takes the adjacency matrix  $\mathbf{A}$  and a graph signal  $\mathbf{f} \in \mathbb{R}^n$  as input. It generates a filtered graph signal  $\mathbf{f}'$  as:

$$\mathbf{f}' = \theta \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{f}, \quad (1)$$

where  $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$  and  $\tilde{\mathbf{D}}$  is the diagonal matrix of  $\tilde{\mathbf{A}}$  with  $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$ . The node representations of all nodes can be viewed as multi-channel graph signals and the  $l$ -th graph convolutional layer with non-linearity is rewritten in a matrix form as:

$$\mathbf{H}^{(l)} = \sigma(\tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-1/2} \mathbf{H}^{(l-1)} \mathbf{W}^{(l)}), \quad (2)$$

where  $\mathbf{H}^{(l)}$  is the output of the  $l$ -th layer,  $\mathbf{H}^{(0)} = \mathbf{X}$ ,  $\mathbf{W}^{(l)}$  is the weight matrix of the  $l$ -th layer and  $\sigma$  is the ReLU activation function.

### 3.1 Laplacian Smoothing in GCN

As shown in Eq. (1), GCN naturally smooths the features in the neighborhoods as each node aggregates information from neighbors. This characteristic is related to Laplacian smoothing and it essentially destroys node similarity in the original feature space. While Laplacian smoothing in GCN has been studied [18, 38], in this work we revisit this process from a new perspective.

Given a graph signal  $\mathbf{f}$  defined on a graph  $\mathcal{G}$  with normalized Laplacian matrix  $\mathbf{L} = \mathbf{I} - \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}$ , the signal smoothness over the graph can be calculated as:

$$\mathbf{f}^T \mathbf{L} \mathbf{f} = \frac{1}{2} \sum_{i,j} \tilde{\mathbf{A}}_{ij} \left( \frac{\mathbf{f}_i}{\sqrt{1+d_i}} - \frac{\mathbf{f}_j}{\sqrt{1+d_j}} \right)^2. \quad (3)$$

where  $d_i$  and  $d_j$  denotes the degree of nodes  $v_i$  and  $v_j$  respectively. Thus a smaller value of  $\mathbf{f}^T \mathbf{L} \mathbf{f}$  indicates smoother graph signal, i.e., smaller signal difference between adjacent nodes.

Suppose that we are given a noisy graph signal  $\mathbf{f}_0 = \mathbf{f}^* + \eta$ , where  $\eta$  is uncorrelated additive Gaussian noise. The original signal  $\mathbf{f}^*$  is assumed to be smooth with respect to the underlying graph  $\mathcal{G}$ . Hence, to recover the original signal, we can adopt the following

objective function:

$$\arg \min_{\mathbf{f}} g(\mathbf{f}) = \|\mathbf{f} - \mathbf{f}_0\|^2 + c\mathbf{f}^T \mathbf{L}\mathbf{f}, \quad (4)$$

Then we have the following lemma.

LEMMA 3.1. *The GCN convolutional operator is one-step gradient descent optimization of Eq. (4) when  $c = \frac{1}{2}$ .*

PROOF. The gradient of  $g(\mathbf{f})$  at  $\mathbf{f}_0$  is calculated as:

$$\nabla g(\mathbf{f}_0) = 2(\mathbf{f}_0 - \mathbf{f}_0) + 2c\mathbf{L}\mathbf{f}_0 = 2c\mathbf{L}\mathbf{f}_0. \quad (5)$$

Then the one step gradient descent at  $\mathbf{f}_0$  with learning rate 1 is formulated as,

$$\begin{aligned} \mathbf{f}_0 - \nabla g(\mathbf{f}_0) &= \mathbf{f}_0 - 2c\mathbf{L}\mathbf{f}_0 \\ &= (\mathbf{I} - 2c\mathbf{L})\mathbf{f}_0 \\ &= (\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}} + \mathbf{L} - 2c\mathbf{L})\mathbf{f}_0. \end{aligned} \quad (6)$$

By setting  $c$  to  $\frac{1}{2}$ , we finally arrives at,

$$\mathbf{f}_0 - \nabla g(\mathbf{f}_0) = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{f}_0. \quad \square$$

By extending the above observation from the signal vector  $\mathbf{f}$  to the feature matrix  $\mathbf{X}$ , we can easily connect the general form of the graph convolutional operator  $\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}$  to Laplacian smoothing. Hence, the graph convolution layer tends to increase the feature similarity between the connected nodes. It is likely to destroy the original feature similarity. In particular, for those disassortative graphs where homophily does not hold, this operation can bring in enormous noise.

### 3.2 Empirical Study for GCN

In the previous subsection, we demonstrated that GCN naturally smooths the features in the neighborhoods, consequently destroying the original feature similarity. In this subsection, we investigate if GCN can preserve node feature similarity via empirical study.

We perform our analysis on two assortative graph datasets (i.e., Cora and Citeseer) and two disassortative graph datasets (i.e., Actor and Cornell). The detailed statistics are shown in Table 2. For each dataset, we first construct two  $k$ -nearest-neighbor ( $k$ NN) graphs based on original features and the hidden representations learned from GCN correspondingly. Together with the original graph, we have, in total, three graphs. Then we calculate the pairwise overlapping for these three graphs. We use  $\mathbf{A}$ ,  $\mathbf{A}_f$  and  $\mathbf{A}_h$  to denote the original graph, the  $k$ NN graph constructed from original features and  $k$ NN graph from hidden representations, respectively. In this experiment,  $k$  is set to 3. Given a pair of adjacency matrices  $\mathbf{A}_1$  and  $\mathbf{A}_2$ , we define the overlapping  $OL(\mathbf{A}_1, \mathbf{A}_2)$  between them as:

$$OL(\mathbf{A}_1, \mathbf{A}_2) = \frac{\|\mathbf{A}_1 \cap \mathbf{A}_2\|}{\|\mathbf{A}_1\|}, \quad (7)$$

where  $\mathbf{A}_1 \cap \mathbf{A}_2$  is the intersection of two matrices and  $\|\mathbf{A}_1\|$  is the number of non-zero elements in  $\mathbf{A}_1$ . In fact,  $OL(\mathbf{A}_1, \mathbf{A}_2)$  indicates the overlapping percentage of edges in the two graphs  $\mathbf{A}_1$  and  $\mathbf{A}_2$ . Larger value of  $OL(\mathbf{A}_1, \mathbf{A}_2)$  means larger overlapping between  $\mathbf{A}_1$  and  $\mathbf{A}_2$ . We show the pairwise overlapping of these three graphs in Table 1. We make the following observations:

**Table 1: Overlapping percentage of different graph pairs.**

Graph Pairs	Cora	Citeseer	Actor	Cornell
$OL(\mathbf{A}_f, \mathbf{A}_h)$	3.15%	3.73%	1.15%	2.35%
$OL(\mathbf{A}_h, \mathbf{A})$	21.24%	18.77%	2.17%	7.74%
$OL(\mathbf{A}_f, \mathbf{A})$	3.88%	3.78%	0.03%	0.91%

- Assortative graphs (Cora and Citeseer) have a higher overlapping percentage of  $(\mathbf{A}_f, \mathbf{A})$  while in disassortative graphs (Actor and Cornell), it is much smaller. It indicates that in assortative graphs the structure information and feature information are highly correlated while in disassortative graphs they are not. It is in line with the definition of assortative and disassortative graphs.
- As indicated by  $OL(\mathbf{A}_f, \mathbf{A})$ , the feature similarity is not very consistent with the graph structure. In assortative graphs, the overlapping percentage of  $(\mathbf{A}_f, \mathbf{A}_h)$  is much lower than that of  $(\mathbf{A}_f, \mathbf{A})$ . GCN will even amplify this inconsistency. In other words, although GCN takes both graph structure and node features as input, it fails to capture more feature similarity information than that the original adjacency matrix has carried.
- By comparing the overlapping percentage of  $(\mathbf{A}_f, \mathbf{A}_h)$  and  $(\mathbf{A}_h, \mathbf{A})$ , we see the overlapping percentage of  $(\mathbf{A}_h, \mathbf{A})$  is consistently higher. Thus, GCN tends to preserve structure similarity instead of feature similarity regardless if the graph is assortative or not.

## 4 THE PROPOSED FRAMEWORK

In this section, we design a node feature similarity preserving graph convolutional framework SimP-GCN by mainly solving the two challenges mentioned in Section 1. An overview of SimP-GCN is shown in Figure 1. It has two major components (i.e., node similarity preserving aggregation and self-supervised learning) corresponding to the two challenges, respectively. The node similarity preserving aggregation component introduces a novel adaptive strategy in Section 4.1 to balance the influence between the graph structure and node features during aggregation. Then the self-supervised learning component in Section 4.2 is to better preserve node similarity by considering both similar and dissimilar pairs. Next we will detail each component.

### 4.1 Node Similarity Preserving Aggregation

In order to endow GCN with the capability of preserving feature similarity, we first propose to integrate the node features with the structure information of the original graph by constructing a new graph that can adaptively balance their influence on the aggregation process. This is achieved by first constructing a  $k$ -nearest-neighbor ( $k$ NN) graph based on the node features and then adaptively integrating it with the original graph into the aggregation process of GCN. Then, learnable self-loops are introduced to adaptively encourage the contribution of a node's own features in the aggregation process.

**4.1.1 Feature Graph Construction.** In feature graph construction, we convert the feature matrix  $\mathbf{X}$  into a feature graph by generating a  $k$ NN graph based on the cosine similarity between the features of each node pair. For a given node pair  $(v_i, v_j)$ , their cosine similarity

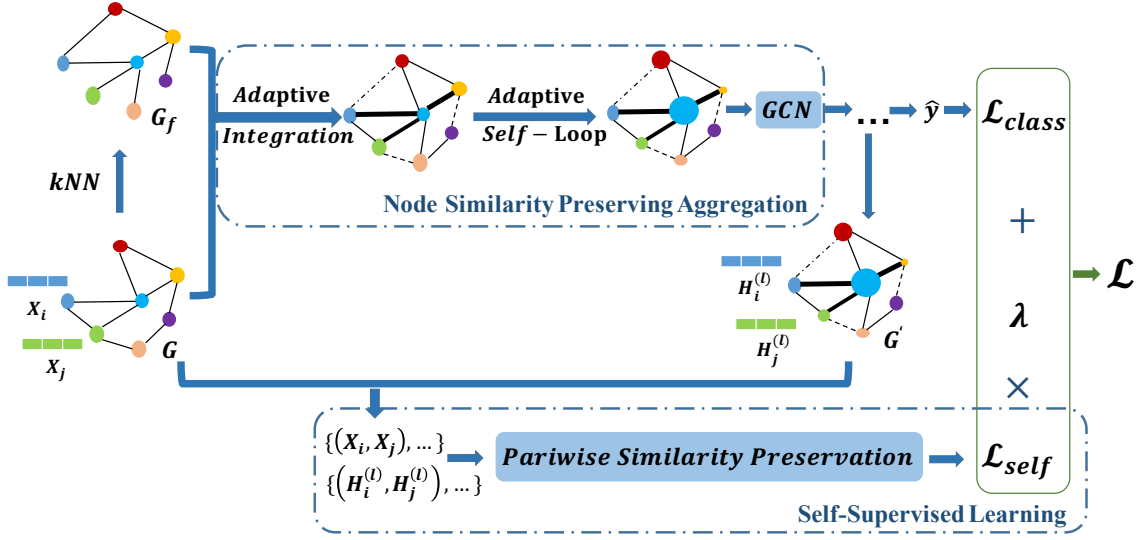


Figure 1: An overall framework of the proposed SimP-GCN.

can be calculated as:

$$S_{ij} = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}. \quad (8)$$

Then we choose  $k = 20$  nearest neighbors following the above cosine similarity for each node and obtain the  $k$ NN graph. We denote the adjacency matrix of the constructed graph as  $\mathbf{A}_f$  and its degree matrix as  $\mathbf{D}_f$ .

**4.1.2 Adaptive Graph Integration.** The graph integration process is adaptive in two folds: (1) each node adaptively balances the information from the original graph and the feature  $k$ NN graph; and (2) each node can adjust the contribution of its node features.

After we obtain the  $k$ NN graph, the propagation matrix  $\mathbf{P}^{(l)}$  in the  $l$ -th layer can be formulated as,

$$\mathbf{P}^{(l)} = \mathbf{s}^{(l)} * \tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-1/2} + (1 - \mathbf{s}^{(l)}) * \mathbf{D}_f^{-1/2} \mathbf{A}_f \mathbf{D}_f^{-1/2}, \quad (9)$$

where  $\mathbf{s}^{(l)} \in \mathbb{R}^n$  is a score vector which balances the effect of the original and feature graphs. Note that “ $\mathbf{a} * \mathbf{M}$ ” denotes the operation to multiply the  $i$ -th element vector  $\mathbf{a}$  with the  $i$ -th row of the matrix  $\mathbf{M}$ . One advantage from  $\mathbf{s}^{(l)}$  is that it allows nodes to adaptively combine information from these two graphs as nodes can have different scores. To reduce the number of parameters in  $\mathbf{s}^{(l)}$ , we model  $\mathbf{s}^{(l)}$  as:

$$\mathbf{s}^{(l)} = \sigma \left( \mathbf{H}^{(l-1)} \mathbf{W}_s^{(l)} + b_s^{(l)} \right), \quad (10)$$

where  $\mathbf{H}^{(l-1)} \in \mathbb{R}^{n \times d^{(l-1)}}$  denotes the input hidden representation from previous layer (with  $\mathbf{H}^{(0)} = \mathbf{X}$ ) and  $\mathbf{W}_s^{(l)} \in \mathbb{R}^{d^{(l-1)} \times 1}$  and  $b_s^{(l)}$  are the parameters for transforming  $\mathbf{H}^{(l-1)}$  into the score vector  $\mathbf{s}^{(l)}$ . Here  $\sigma$  denotes an activation function and we apply the sigmoid function. In this way, the number of parameters for construing  $\mathbf{s}^{(l)}$  has reduced from  $n$  to  $d^{(l-1)} + 1$ .

**4.1.3 Adaptive Learnable Self-loops.** In the original GCN, given a node  $v_i$ , one self-loop is added to include its own features into

the aggregation process. Thus, if we want to preserve more information from its original features, we can add more self-loops to  $v_i$ . However, the importance of node features could be distinct for different nodes; and different numbers of self-loops are desired for nodes. Thus, we propose to add a learnable diagonal matrix  $\mathbf{D}_K^{(l)} = \text{diag}(K_1^{(l)}, K_2^{(l)}, \dots, K_n^{(l)})$  to the propagation matrix  $\mathbf{P}^{(l)}$ ,

$$\tilde{\mathbf{P}}^{(l)} = \mathbf{P}^{(l)} + \gamma \mathbf{D}_K^{(l)}, \quad (11)$$

where  $\mathbf{D}_K^{(l)}$  adds learnable self-loops to the integrated graph. In particular,  $K_i^{(l)}$  indicates the number of self-loops added to node  $v_i$  at the  $l$ -th layer.  $\gamma$  is a predefined hyper-parameter to control the contribution of self-loops. To reduce the number of parameters, we use a linear layer to learn  $K_1^{(l)}, K_2^{(l)}, \dots, K_n^{(l)}$  as:

$$K_i^{(l)} = \mathbf{H}_i^{(l-1)} \mathbf{W}_K^{(l)} + b_K^{(l)}, \quad (12)$$

where  $\mathbf{H}_i^{(l-1)}$  is the hidden presentation of  $v_i$  at the layer  $l - 1$ ;  $\mathbf{W}_K^{(l)} \in \mathbb{R}^{d^{(l-1)} \times 1}$  and  $b_K^{(l)}$  are the parameters for learning the self-loops  $K_1^{(l)}, K_2^{(l)}, \dots, K_n^{(l)}$ .

**4.1.4 The Classification Loss.** Ultimately the hidden representations can be formulated as,

$$\mathbf{H}^{(l)} = \sigma(\tilde{\mathbf{P}}^{(l)} \mathbf{H}^{(l-1)} \mathbf{W}^{(l)}), \quad (13)$$

where  $\mathbf{W}^{(l)} \in \mathbb{R}^{d^{(l-1)} \times d^{(l)}}$ ,  $\sigma$  denotes an activation function such as the ReLU function. We denote the output of the last layer as  $\hat{\mathbf{H}}$ , then the classification loss is shown as,

$$\mathcal{L}_{class} = \frac{1}{|\mathcal{D}_L|} \sum_{(v_i, y_i) \in \mathcal{D}_L} \ell(\text{softmax}(\hat{\mathbf{H}}_i), y_i), \quad (14)$$

where  $\mathcal{D}_L$  is the set of labeled nodes,  $y_i$  is the label of node  $v_i$  and  $\ell(\cdot, \cdot)$  is the loss function to measure the difference between predictions and true labels such as cross entropy.

## 4.2 Self-Supervised Learning

Although the constructed  $k$ NN feature graph component in Eq. (10) plays a critical role of pushing nodes with similar features to become similar, it does not directly model feature dissimilarity. In other words, it only can effectively preserve the top similar pairs of nodes to have similar representations and will not push nodes with dissimilar original features away from each other in the learned embedding space. Thus, to better preserve pairwise node similarity, we incorporate the complex pairwise node relations by proposing a contrastive self-supervised learning component to SimP-GCN.

Recently, self-supervised learning techniques have been applied to graph neural networks for leveraging the rich information in tremendous unlabeled nodes [11, 12, 39, 40]. Self-supervised learning first designs a domain specific pretext task to assign constructed labels for nodes and then trains the model on the pretext task to learn better node representations. Following the joint training manner described in [12, 40], we design a contrastive pretext task where the self-supervised component is asked to predict pairwise feature similarity. In detail, we first calculate pairwise similarity for each node pair and sample node pairs to generate self-supervised training samples. Specifically, for each node, we sample its  $m$  most similar nodes and  $m$  most dissimilar nodes. Then the pretext task can be formulated as a regression problem and the self-supervised loss can be stated as,

$$\mathcal{L}_{self}(\mathbf{A}, \mathbf{X}) = \frac{1}{|\mathcal{T}|} \sum_{(v_i, v_j) \in \mathcal{T}} \|f_w(\mathbf{H}_i^{(l)} - \mathbf{H}_j^{(l)}) - S_{ij}\|^2, \quad (15)$$

where  $\mathcal{T}$  is the set of sampled node pairs,  $f_w$  is a linear mapping function,  $S_{ij}$  is defined in Eq. (8), and  $\mathbf{H}_i^{(l)}$  is the hidden representation of node  $v_i$  at  $l$ -th layer. Note that we can formulate the pretext task as a classification problem to predict whether the given node pair is similar or not (or even the levels of similarity/dissimilarity). However, in this work we adopt the regression setting and use the first layer’s hidden representation, i.e.,  $l = 1$  while setting  $m = 5$ .

## 4.3 Objective Function and Complexity Analysis

With the major components of SimP-GCN, next we first present the overall objective function where we jointly optimize the classification loss along with the self-supervised loss. Thereafter, we present a complexity analysis and discussion on the proposed framework.

**4.3.1 Overall Objective Function.** The overall objective function can be stated as,

$$\begin{aligned} \min \mathcal{L} &= \mathcal{L}_{class} + \lambda \mathcal{L}_{self} \\ &= \frac{1}{|\mathcal{D}_L|} \sum_{(v_i, y_i) \in \mathcal{D}_L} \ell(\text{softmax}(\hat{\mathbf{H}}_i), y_i) \\ &\quad + \frac{\lambda}{|\mathcal{T}|} \sum_{(v_i, v_j) \in \mathcal{T}} \|f_w(\mathbf{H}_i^{(1)} - \mathbf{H}_j^{(1)}) - S_{ij}\|^2 \end{aligned} \quad (16)$$

where  $\lambda$  is a hyper-parameter that controls the contribution of self-supervised loss (i.e.,  $\mathcal{L}_{self}$  of Eq. (15)) in addition to the traditional classification loss (i.e.,  $\mathcal{L}_{class}$  of Eq. (14)).

**Table 2: Dataset Statistics.**

Datasets	Assortative			Disassortative			
	Cora	Cite.	Pubm.	Actor	Corn.	Texas	Wisc.
#Nodes	2,708	3,327	19,717	7,600	183	183	251
#Edges	5,429	4,732	44,338	33,544	295	309	499
#Features	1,433	3,703	500	931	1,703	1,703	1,703
#Classes	7	6	3	5	5	5	5

**4.3.2 Complexity Analysis.** Here, we compare the proposed method, SimP-GCN, with vanilla GCN by analyzing the additional complexity in terms of time and model parameters.

**Time Complexity.** In comparison to vanilla GCN, the additional computational requirement mostly comes from calculating the pairwise feature similarity for the  $k$ NN graph construction and self-supervised component. Its time complexity is  $O(n^2)$  if done naïvely. However, calculating pairwise similarity can be made significantly more efficient. For example, the calculations are inherently parallelizable. In addition, this has been a well studied problem and in the literature there exist approximation methods that could be used to significantly speedup the computation for larger graphs, such as [4] that presented a divide-and-conquer method via Recursive Lanczos Bisection, or [7] that is inherently suitable for MapReduce where they empirically achieved approximate  $k$ NN graphs in  $O(n^{1.14})$ .

**Model Complexity.** As shown in Section 4.1 and Section 4.2, our model introduces additional parameters  $\mathbf{W}_s^{(l)}, b_s^{(l)}, \mathbf{W}_K^{(l)}, b_K^{(l)}$  and linear mapping  $f_w$ . It should be noted that  $\mathbf{W}_s^{(l)}, \mathbf{W}_K^{(l)} \in \mathbb{R}^{d^{(l-1)} \times 1}$  and  $b_s^{(l)}, b_K^{(l)} \in \mathbb{R}$ . Since  $f_w$  transforms  $\mathbf{H}$  to a one-dimensional vector, the weight matrix in  $f_w$  has a shape of  $d^{(l)} \times 1$ . Hence, compared with GCN, the overall additional parameters of our model are  $O(d^{(l)})$  where  $d^{(l)}$  is the input feature dimension in the  $l$ -th layer. It suggests that our model only introduces additional parameters that is linear to the feature dimension.

## 5 EXPERIMENT

In this section, we evaluate the effectiveness of SimP-GCN under different settings. In particular, we aim to answer the following questions:

- **RQ1:** How does SimP-GCN perform on both assortative and disassortative graphs?
- **RQ2:** Can SimP-GCN boost model robustness and defend against graph adversarial attacks?
- **RQ3:** Does the framework work as designed? and How do different components affect the performance of SimP-GCN?

Note that though there are many scenarios where node similarity is important, we focus on the scenarios of disassortative graphs and graphs manipulated by adversarial attacks to illustrate the advantages of the proposed framework by preserving node similarity.

### 5.1 Experimental Settings

**5.1.1 Datasets.** Since the performance of graph neural networks can be significantly different on assortative and disassortative graphs, we select several representative datasets from both categories to conduct the experiments. Specifically, for assortative graphs we

adopt three citation networks that are commonly used in the GNN literature, i.e., Cora, Citeseer and Pubmed [15]. For disassortative graphs, we use one actor co-occurrence network, i.e., Actor [31], and three webpage datasets, i.e., Cornell, Texas and Wisconsin [27]. The statistics of these datasets are shown in Table 2.

**5.1.2 Baselines.** To evaluate the effectiveness of the proposed framework, we choose the following representative semi-supervised learning baselines including the state-of-the-art GNN models:

- **LP** [43]: Label Propagation (LP) explores structure and label information by a Gaussian random field model. Note that LP does not exploit node features.
- **GCN** [15]: GCN is one of the most popular graph convolutional models and our proposed model is based on it.
- **$k$ NN-GCN** [8]: It is a variant of GCN which takes the  $k$ -nearest neighbor graph created from the node features as input. Here we use it as a baseline to check the performance of directly employing the  $k$ NN graph without using the original graph.
- **(A+kNN)-GCN**: GCN that takes  $A + A_f$  as the input graph.
- **GAT** [33]: Graph attention network (GAT) employs attention mechanism and learns different scores to different nodes within the neighborhood. It is widely used as a GNN baseline.
- **JK-Net** [37]: JK-Net uses dense connections to leverage different neighbor ranges for each node to learn better representations.
- **GCNII** [5]: Based on GCN, GCNII employs residual connection and identity mapping to achieve better performance. GCNII\* is a variant of GCNII with more parameters.
- **Geom-GCN** [27]: Geom-GCN explores to capture long-range dependencies in disassortative graphs. It uses the geometric relationships defined in the latent space to build structural neighborhoods for aggregation. There are three variants of Geom-GCN: Geom-GCN-I, Geom-GCN-P and Geom-GCN-S.

Note that **Geom-GCN** is mainly designed for disassortative graphs; thus we only report its performance on disassortative graphs.

**5.1.3 Parameter Setting.** For experiments on assortative graphs, we set the number of layers in GCN and SimP-GCN to 2, with hidden units 128,  $L_2$  regularization  $5e-4$ , dropout rate 0.5, epochs 200 and learning rate 0.01. In SimP-GCN, the weighting parameter  $\lambda$  is searched from  $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ ,  $\gamma$  is searched from  $\{0.01, 0.1\}$  and the initialization of  $b_s$  is searched from  $\{0, 2\}$ . Since GCNII and JK-Net with multi-layers can use much higher memory due to more parameters, we restrict the depth to 4. For experiments on disassortative graphs, we set the number of layers in GCN and SimP-GCN to 2, with learning rate 0.05, dropout rate 0.5, epochs 500 and patience 100. The number of hidden units is tuned from  $\{16, 32, 48\}$  and  $L_2$  regularization is tuned from  $\{5e-4, 5e-5\}$ . SimP-GCN further searches  $\lambda$  from  $\{0.1, 1, 10\}$  and  $\gamma$  from  $\{0.01, 0.1, 1\}$ . For other baseline methods, we use the default parameter settings in the authors' implementation.

## 5.2 Performance Comparison

In this subsection, we answer the first question and compare the performance on assortative and disassortative graphs, respectively.

**5.2.1 Performance Comparison on Assortative Graphs.** For the experiments on assortative graphs, we follow the widely used semi-supervised setting in [15] with 20 nodes per class for training, 500

**Table 3: Node classification accuracy (%) on assortative graphs. The best performance is highlighted in bold.**

Method	Cora	Citeseer	Pubmed
LP	74.6	57.7	71.6
GCN	81.3	71.5	79.3
$k$ NN-GCN	67.4	68.7	78.9
(A+kNN)-GCN	79.1	71.1	80.8
GAT	<b>83.1</b>	70.8	78.5
JK-Net	80.3	68.5	78.3
GCNII	82.6	68.9	78.8
GCNII*	82.3	67.9	78.2
SimP-GCN	82.8	<b>72.6</b>	<b>81.1</b>

nodes for validation and 1000 nodes for testing. We report the average accuracy of 10 runs in Table 3. On the three assortative graphs, SimP-GCN consistently improves GCN and achieves the best performance in most settings. The improvement made by SimP-GCN demonstrates that preserving feature similarity can benefit GCN in exploiting structure and feature information. It is worthwhile to note that LP and  $k$ NN, which only take the graph structure or node features as input, can achieve reasonable performance, but they cannot beat GCN. The variant (A+kNN)-GCN cannot outperform  $k$ NN-GCN or GCN. This suggests that learning a balance between graph structure and node features is of great significance.

**Table 4: Node classification accuracy (%) on disassortative graphs. The best performance is highlighted in bold.**

Method	Actor	Cornell	Texas	Wisconsin
LP	23.53	35.95	32.70	29.41
GCN	26.86	52.70	52.16	45.88
$k$ NN-GCN	33.77	81.35	79.73	81.35
(A+kNN)-GCN	32.83	76.49	79.60	80.98
GAT	28.45	54.32	58.38	49.41
JK-Net	27.41	57.84	55.95	50.78
Geom-GCN-I	29.09	56.76	57.58	58.24
Geom-GCN-P	31.63	60.81	67.57	64.12
Geom-GCN-S	30.30	55.68	59.73	56.67
GCNII	33.91	74.86	69.46	74.12
GCNII*	35.18	76.49	77.84	81.57
SimP-GCN	<b>36.20</b>	<b>84.05</b>	<b>81.62</b>	<b>85.49</b>

**5.2.2 Performance Comparison on Disassortative Graphs.** In this experiment, we report the performance on four disassortative graphs, i.e., Actor, Cornell, Texas and Wisconsin. Following the common setting of experiments on disassortative graphs [23, 27], we randomly split nodes of each class into 60%, 20%, and 20% for training, validation and test.

Note that both Geom-GCN and GCNII are the recent models that achieve the state-of-the-art performance on disassortative graphs. We report the average accuracy of all models on the test sets over 10 random splits in Table 4. We reuse the metrics reported in [5] for GCN, GAT and Geom-GCN. From the results we can

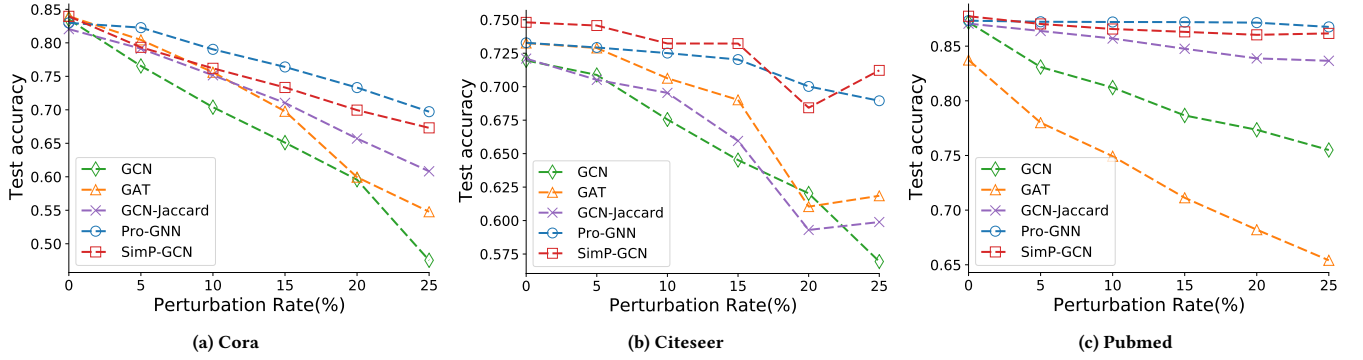


Figure 2: Node classification accuracy under non-targeted attack (*metattack*).

see that the performance of LP is extremely low on all datasets, which indicates that structure information is not very useful for the downstream task. Besides, the fact that GCN, GAT, JK-Net and Geom-GCN cannot outperform  $k$ NN-GCN verifies that feature information in those disassortative graphs are much more important and the original structure information could even be harmful. We note that  $(A+k$ NN)-GCN cannot improve  $k$ NN-GCN. It supports that simply combining structure and features is not sufficient. SimP-GCN consistently improves GCN by a large margin and achieves state-of-the-art results on four disassortative graphs.

### 5.3 Adversarial Robustness

Traditional deep neural networks can be easily fooled by adversarial attacks [19, 20, 36]. Similarly, recent research has also demonstrated that graph neural networks are vulnerable to adversarial attacks [14, 30, 44]. In other words, unnoticeable perturbation on graph structure can significantly reduce their performances. Such vulnerability has raised great concern for applying graph neural networks in safety-critical applications. Although attackers can perturb both node features and graph structures, the majority of existing attacks focus on changing graph structures. Recall that SimP-GCN is shown to greatly improve GCN when the structure information is not very useful in Section 5.2.2. Therefore, in this subsection, we are interested in examining its potential benefit on adversarial robustness. Specifically, we evaluate the node classification performance of SimP-GCN under non-targeted graph adversarial attacks. The goal of non-targeted attack is to degrade the overall performance of GNNs on the whole node set. We adopt *metattack* [45] as the attack method. We focus on attacking graph structure and vary the perturbation rate, i.e., the ratio of changed edges, from 0 to 25% with a step of 5%. To make better comparison, we include GCN, GAT and the state-of-the-art defense models, GCN-Jaccard [34] and Pro-GNN [14] implemented by DeepRobust<sup>1</sup> [21], as baselines and use the default hyper-parameter settings in the authors' implementations. For each dataset, the hyper-parameters of SimP-GCN are set as the same values under different perturbation rates. We follow the parameter settings of baseline and attack methods in [14] and evaluate models' robustness on Cora, Citeseer and Pubmed datasets. Furthermore, as suggested in [14], we randomly

choose 10% of nodes for training, 10% of nodes for validation and 80% of nodes for test. All experiments are repeated 10 times and the average accuracy of different methods under various settings is shown in Figure 2.

As we can observe from Figure 2, SimP-GCN consistently improves the performance of GCN under different perturbation rates of adversarial attack on all three datasets. Specifically, SimP-GCN improves GCN by a larger margin when the perturbation rate is higher. For example, it achieves over 20% improvement over GCN under the 25% perturbation rate on Cora dataset. This is because the structure information will be very noisy and misleading when the graph is heavily poisoned. In this case, node feature similarity can provide strong guidance to train the model and boost model robustness. SimP-GCN always outperforms GCN-Jaccard and shows comparable performance to Pro-GNN on all datasets. It even obtains the best performance on the Citeseer dataset. These observations suggest that by preserving node similarity, SimP-GCN can be robust to existing adversarial attacks.

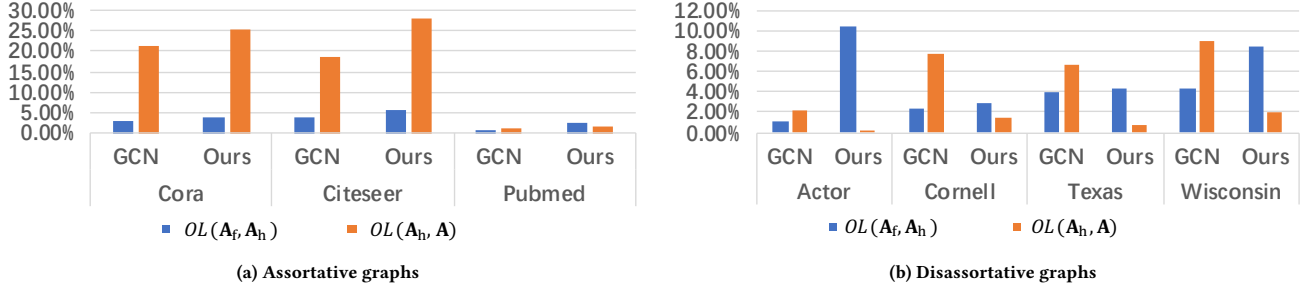
### 5.4 Further Probe

In this subsection, we take a deeper examination on the proposed framework to understand how it works and how each component affects its performance.

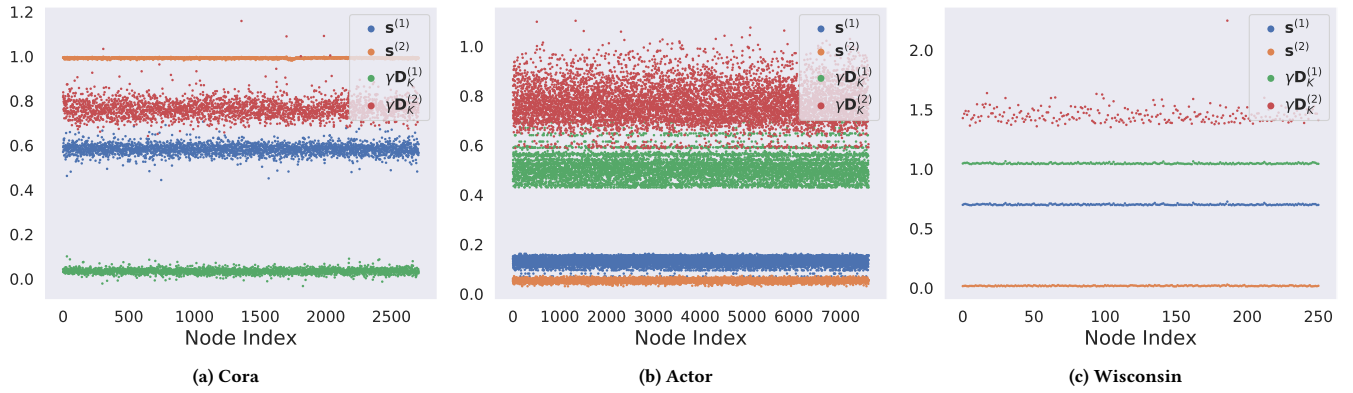
**5.4.1 Is Node Similarity Well Preserved?** Next, we investigate whether SimP-GCN can preserve node feature similarity. Following the same evaluation strategy in Section 3.2, we focus on the overlapping percentage of graphs pairs. Specifically, we construct  $k$ NN graphs from the learned hidden representations of GCN and SimP-GCN, denoted as  $A_h$ , and calculate the overlapping percentages  $OL(A_f, A_h)$  and  $OL(A_h, A)$  on all datasets. The results are shown in Figure 3. From the results, we make the following observations:

- For both assortative and disassortative graphs, SimP-GCN always improves the overlapping percentage between feature and hidden graphs, i.e.,  $OL(A_f, A_h)$ . It shows that SimP-GCN can effectively preserve more information from feature similarity. Especially in disassortative graphs where homophily is low, SimP-GCN essentially boosts  $OL(A_f, A_h)$  by a large margin, e.g., roughly 10% on the Actor dataset.
- In disassortative graphs, SimP-GCN decreases the overlapping percentage between hidden and original graphs, i.e.,  $OL(A_h, A)$ .

<sup>1</sup><https://github.com/DSE-MSU/DeepRobust>



**Figure 3: Overlapping percentage on assortative and disassortative graphs. “Ours” means SimP-GCN. Blue bar indicates the overlapping between feature and hidden graphs; and orange bar denotes the overlapping between hidden and original graphs.**



**Figure 4: Learned values of  $s^{(1)}$ ,  $s^{(2)}$ ,  $D_n^{(1)}$  and  $D_n^{(2)}$  on Cora, Actor and Wisconsin.**

One reason is that the structure information in disassortative graphs is less useful (or even harmful) to the downstream task. Such phenomenon is also in line with our observation in Section 5.2.2. On the contrary, in assortative graphs both  $OL(A_h, A)$  and  $OL(A_f, A_h)$  are improved, which indicates that SimP-GCN can explore more information from both structure and features for assortative graphs.

**5.4.2 Do the Score Vectors and Self-loop Matrices Work?** To study how the score vectors ( $s^{(1)}$ ,  $s^{(2)}$ ) and learnable self-loop matrices ( $D_n^{(1)}$ ,  $D_n^{(2)}$ ) in Eq. (10) benefit SimP-GCN, we set  $\gamma$  to 0.1 and visualize the learned values in Figure 4. Due to the page limit, we only report the results on Cora, Actor and Wisconsin. From the results, we have the following findings:

- The learned values vary in a range, indicating that the designed aggregation scheme can adapt the information differently for individual nodes. For example, in Cora dataset,  $s^{(1)}$ ,  $s^{(2)}$ ,  $\gamma D_n^{(1)}$  and  $\gamma D_n^{(2)}$  are in the range of  $[0.44, 0.71]$ ,  $[0.98, 1.00]$ ,  $[-0.03, 0.1]$  and  $[0.64, 1.16]$ , respectively.
- On the assortative graph (Cora),  $\gamma D_n^{(1)}$  is extremely small and the model is mainly balancing the information from original and feature graphs. On the contrary, in disassortative graphs (Actor and Wisconsin),  $\gamma D_n^{(1)}$  has much larger values, indicating that node original features play a significant role in making predictions for disassortative graphs.

**5.4.3 Ablation Study.** To get a better understanding of how different components affect the model performance, we conduct ablation studies and answer the third question in this subsection. Specifically, we build the following ablations:

- Keeping self-supervised learning (SSL) component and all other components. This one is our original proposed model.
- Keeping SSL component but removing the learnable diagonal matrix  $D_n$  from the propagation matrix  $\tilde{P}$ , i.e., setting  $\gamma$  to 0.
- Removing SSL component.
- Removing SSL component and the learnable diagonal matrix  $D_n$ .

Since SimP-GCN achieves the most significant improvement on disassortative graphs, we only report the performance on disassortative graphs. We use the best performing hyper-parameters found for the results in Table 4 and report the average accuracy of 10 runs in Table 5. By comparing the ablations with GCN, we observe that all components contribute to the performance gain:  $A_f$  and  $D_n$  essentially boost the performance while the SSL component can further improve the performance based on  $A_f$  and  $D_n$ .

## 6 CONCLUSION

Graph neural networks extract effective node representations by aggregating and transforming node features within the neighborhood. We show that the aggregation process inevitably breaks node similarity of the original feature space through theoretical and empirical analysis. Based on these findings, we introduce a node

**Table 5: Ablation study results (% test accuracy) on disassortative graphs.**

Ablation	Actor	Corn.	Texa.	Wisc.
SimP-GCN				
- with SSL and $(A, A_f, D_n)$	36.20	84.05	81.62	85.49
- with SSL and $(A, A_f)$	35.75	65.95	71.62	81.37
- w/o SSL and with $(A, A_f, D_n)$	36.09	82.70	79.73	84.12
- w/o SSL and with $(A, A_f)$	34.68	64.59	68.92	81.57
GCN	26.86	52.70	52.16	45.88

similarity preserving graph convolutional neural network model, SimP-GCN, which effectively and adaptively balances the structure and feature information as well as captures pairwise node similarity via self-supervised learning. Extensive experiments demonstrate that SimP-GCN outperforms representative baselines on a wide range of real-world datasets. As one future work, we plan to explore the potential of exploring node similarity in other scenarios such as the over-smoothing issue and low-degree nodes.

## 7 ACKNOWLEDGEMENTS

This research is supported by the National Science Foundation (NSF) under grant numbers CNS1815636, IIS1928278, IIS1714741, IIS1845081, IIS1907704 and IIS1955285.

## REFERENCES

- [1] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).
- [2] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013).
- [3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203* (2013).
- [4] Jie Chen, Haw-ren Fang, and Yousef Saad. 2009. Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection. *Journal of Machine Learning Research* 10, 9 (2009).
- [5] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. 2020. Simple and Deep Graph Convolutional Networks. *arXiv preprint arXiv:2007.02133* (2020).
- [6] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*.
- [7] Wei Dong, Charikar Moses, and Kai Li. 2011. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*. 577–586.
- [8] Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. 2019. Learning discrete structures for graph neural networks. *ICML* (2019).
- [9] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.
- [10] Will Hamilton, Zitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*.
- [11] Kaveh Hassani and Amir Hosein Khasahmadi. 2020. Contrastive Multi-View Representation Learning on Graphs. *arXiv preprint arXiv:2006.05582* (2020).
- [12] Wei Jin, Tyler Derr, Haochen Liu, Yiqi Wang, Suhang Wang, Zitao Liu, and Jiliang Tang. 2020. Self-supervised Learning on Graphs: Deep Insights and New Direction. *arXiv:cs.LG/2006.10141*
- [13] Wei Jin, Yaxin Li, Han Xu, Yiqi Wang, and Jiliang Tang. 2020. Adversarial Attacks and Defenses on Graphs: A Review and Empirical Study. *arXiv preprint arXiv:2003.00653* (2020).
- [14] Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. 2020. Graph Structure Learning for Robust Graph Neural Networks. *arXiv preprint arXiv:2005.10203* (2020).
- [15] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [16] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).
- [17] Hang Li, Haozheng Wang, Zhenglu Yang, and Haochen Liu. 2017. Effective representing of information network by variational autoencoder. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2103–2109.
- [18] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. *arXiv preprint arXiv:1801.07606* (2018).
- [19] Yingwei Li, Song Bai, Cihang Xie, Zhenyu Liao, Xiaohui Shen, and Alan Yuille. 2019. Regional Homogeneity: Towards Learning Transferable Universal Adversarial Perturbations Against Defenses. *arXiv preprint arXiv:1904.00979* (2019).
- [20] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan L Yuille. 2020. Learning Transferable Adversarial Examples via Ghost Networks.. In *AAAI*.
- [21] Yaxin Li, Wei Jin, Han Xu, and Jiliang Tang. 2020. DeepRobust: A PyTorch Library for Adversarial Attacks and Defenses. *arXiv preprint arXiv:2005.06149* (2020).
- [22] Meng Liu, Hongyang Gao, and Shuiwang Ji. 2020. Towards Deeper Graph Neural Networks. *arXiv preprint arXiv:2007.09296* (2020).
- [23] Meng Liu, Zhengyang Wang, and Shuiwang Ji. 2020. Non-Local Graph Neural Networks. *arXiv preprint arXiv:2005.14612* (2020).
- [24] Yao Ma and Jiliang Tang. 2020. *Deep Learning on Graphs*. Cambridge University Press.
- [25] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.
- [26] Mark EJ Newman. 2002. Assortative mixing in networks. *Physical review letters* 89, 20 (2002), 208701.
- [27] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. 2020. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287* (2020).
- [28] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. 2019. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*.
- [29] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine* 30, 3 (2013), 83–98.
- [30] Yiwei Sun, Suhang Wang, Xianfeng Tang, Tsung-Yu Hsieh, and Vasant Honavar. 2020. Adversarial Attacks on Graph Neural Networks via Node Injections: A Hierarchical Reinforcement Learning Approach. In *WWW*.
- [31] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. 2009. Social influence analysis in large-scale networks. In *KDD*.
- [32] Xianfeng Tang, Huaxiu Yao, Yiwei Sun, Yiqi Wang, Jiliang Tang, Charu Aggarwal, Prasenjit Mitra, and Suhang Wang. 2020. Graph Convolutional Networks against Degree-Related Biases. *arXiv preprint arXiv:2006.15643* (2020).
- [33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [34] Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. 2019. Adversarial examples on graph data: Deep insights into attack and defense. *arXiv preprint arXiv:1903.01610* (2019).
- [35] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2019. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596* (2019).
- [36] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil Jain. 2019. Adversarial attacks and defenses in images, graphs and text: A review. *arXiv preprint arXiv:1909.08072* (2019).
- [37] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2018. Representation learning on graphs with jumping knowledge networks. *arXiv preprint arXiv:1806.03536* (2018).
- [38] Chaoqi Yang, Ruijie Wang, Shuochao Yao, Shengzhong Liu, and Tarek Abdelzaher. 2020. Revisiting “Over-smoothing” in Deep GCNs. *arXiv preprint arXiv:2003.13663* (2020).
- [39] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph Contrastive Learning with Augmentations. *arXiv preprint arXiv:2010.13902* (2020).
- [40] Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. 2020. When Does Self-Supervision Help Graph Convolutional Networks? *ICML* (2020).
- [41] Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*. 5165–5175.
- [42] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434* (2018).
- [43] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*.
- [44] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. 2018. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.
- [45] Daniel Zügner and Stephan Günnemann. 2019. Adversarial attacks on graph neural networks via meta learning. *arXiv preprint arXiv:1902.08412* (2019).