Heterophilous Distribution Propagation for Graph Neural Networks

Zhuonan Zheng[a,b], Sheng Zhou[b,c,*], Hongjia Xu[a,b], Ming Gu[a,b], Yilun Xu[a,b], Ao Li[d], Yuhong Li[d], Jingjun Gu[a,b], Jiajun Bu[a,b]

[a]*College of Computer Science, Zhejiang University, Hangzhou, 310027, China*
[b]*Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, Zhejiang University, Hangzhou, 310027, China*
[c]*School of Software Technology, Zhejiang University, Ningbo, 315048, China*
[d]*Alibaba Group, Hangzhou, 310052, China*

Abstract

Graph Neural Networks (GNNs) have achieved remarkable success in various graph mining tasks by aggregating information from neighborhoods for representation learning. The success relies on the homophily assumption that nearby nodes exhibit similar behaviors, while it may be violated in many real-world graphs. Recently, heterophilous graph neural networks (HeterGNNs) have attracted increasing attention by modifying the neural message passing schema for heterophilous neighborhoods. However, they suffer from insufficient **neighborhood partition** and **heterophily modeling**, both of which are critical but challenging to break through. To tackle these challenges, in this paper, we propose heterophilous distribution propagation (HDP) for graph neural networks. Instead of aggregating information from all neighborhoods, HDP adaptively separates the neighbors into homophilous and heterphilous parts based on the pseudo assignments during training. The heterophilous neighborhood distribution is learned with orthogonality-oriented constraint via a trusted prototype contrastive learning paradigm. Both the homophilous and heterophilous patterns are propagated with a novel semantic-aware message-passing mechanism. We conduct extensive experiments on 9 benchmark datasets with different levels of homophily. Experimental results show that our method outperforms representative baselines on heterophilous datasets.

Keywords: Graph Neural Networks, Graph Representation Learning, Graph Heterophily

*Corresponding author

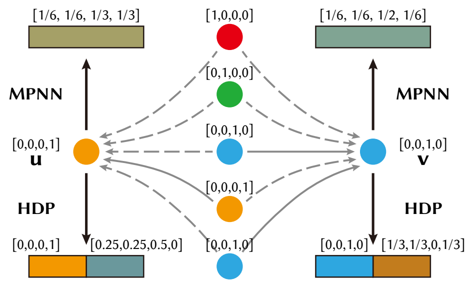
Email addresses: zhengzn@zju.edu.cn (Zhuonan Zheng), zhousheng_zju@zju.edu.cn (Sheng Zhou), xu_hj@zju.edu.cn (Hongjia Xu), guming444@zju.edu.cn (Ming Gu), yilun.xu@zju.edu.cn (Yilun Xu), jianzhen.la@alibaba-inc.com (Ao Li), daniel.lyh@alibaba-inc.com (Yuhong Li), gjj@zju.edu.cn (Jingjun Gu), bjj@zju.edu.cn (Jiajun Bu)

1. Introduction

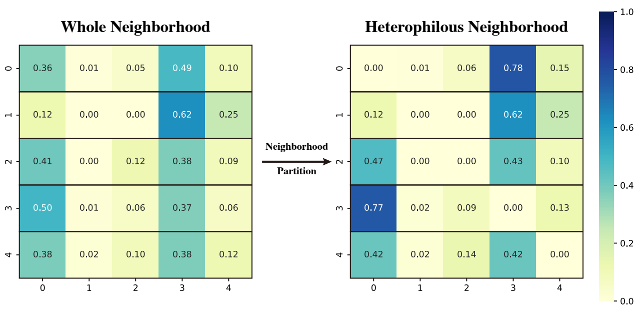
Graph Neural Networks (GNNs) aim at learning effective representations for graph data, which have demonstrated exceptional performance across a range of graph mining tasks, including node classification [1, 2], link prediction [3, 4], graph classification [5, 6], and anomaly detection [7, 8]. The majority of existing GNNs utilize the neural message passing (NMP) schema and aggregate information from the neighborhood for representation learning. This is predicated on the homophily assumption, which suggests that nodes in close proximity within a graph are likely to exhibit similar behaviors, such as labels and features. This assumption has been widely observed in bibliographic graphs [9] and online social networks [10, 11]. However, many types of graphs challenge the homophily assumption that connected nodes can exhibit **heterophily** patterns, which presents significant challenges for the application of GNNs.

To tackle this challenge, Heterophilous Graph Neural Networks (HeterGNNs) [12, 13, 14, 15] have attracted increasing attention from both academic and industry communities in the past few years. Specifically, they have mainly focused on modifying the message-passing process by targeting the characteristics of heterophilous graphs from different perspectives. Among them, early methods primarily adjust the **scope of message passing**, such as decreasing the proportion of heterophilous neighbors by enlarging the high-order neighborhood [16]. Other methods alter the **message passing process itself**, for instance, by assigning varying weights to neighbors based on the similarity of their representations [17]. Further, some methods modify the **update process**, such as separating the representations of ego nodes and their neighboring nodes [13].

Despite their achievements, we argue that existing methods still suffer from the following shortcomings: (i) Insufficient **neighborhood partition**. Unlike the homophilous neighborhood that exhibits consistent patterns, the heterophilous neighborhood introduces significantly different patterns and deserves specific modeling. However, most existing methods either do not distinguish between homophilous and heterophilous neighborhoods, or simply utilize fixed thresholds for naive partition. This not only obscures the unique characteristics of the corresponding class which is shared by homophilous neighbors, but also limits the role of heterophilous neighbors, as we show in Figure 1a. (ii) Insufficient **heterophily modeling**. As previously mentioned, many existing HeterGNNs aggregated information from both homophilous and heterophilous neighbors at the same time without individual heterophily modeling. However, given that heterophilous neighbors originate from multiple categories, simply aggregating and fusing their messages with central nodes



(a) A toy example of neighborhood partition



(b) Observed connection preference in Cornell.

Figure 1: The illustration of neighborhood partition and heterophily modeling. (a) Nodes **u** and **v** belong to different classes but have identical neighborhoods, where colors denote classes. Without neighborhood partition, message-passing neural networks(MPNN) produce confused node representations due to similar message sources. On the contrary, partitioning neighborhoods explicitly and handling them separately can increase the discriminability of representations. (b) We visualize the observed connection preference among classes in Cornell, where each row denotes the connect probabilities between the corresponding class and others. In the whole neighborhood, the connection preferences of different classes are similar. After neighborhood partition, the connection preferences in heterophilous neighborhoods show the discriminability. Thus, heterophily modeling can also bring additional information for discriminative representation learning beyond homophilous neighbors.

and homophilous neighbors will lose critical heterophilous connection patterns. Figure 1b shows the advantages of neighborhood partition and individual heterophily modeling for real-world datasets.

Although important, overcoming the aforementioned shortcomings is quite challenging: Firstly, the accurate partition of homophilous and heterphilous neighborhoods relies on the node labels. Given the limited number of semi-supervised labels, we can only rely on the pseudo labels produced by the HeterGNNs. The **mutually dependent** between effective representation learning and neighborhood partition poses significant challenges. Secondly, heterophily modeling relies on the diverse connections between nodes from different classes. However, due to the network sparsity and scarcity of node labels, we can not model the heterophily connection patterns for each node from its limited neighborhood. Instead, the patterns are expected to **propagate along with the homophilous edges** for more comprehensive modeling.

To tackle the above challenges, in this paper, we propose **H**eterophilous **D**istribution **P**ropagation for Graph Neural Networks (HDP). More specifically, given a heterophilous graph with an unknown heterophily ratio, HDP first estimates the heterophily level and partition the neighborhood according to the predicted semantic assignments, which are dynamically updated along with the training process. On this basis, we model the heterophilous neighborhood distribution via a simple but effective operator with orthogonality-oriented constraint as a trusted prototype contrastive learning paradigm. We also theoretically prove the advantages of the operator. To further enhance the heterophily modeling, we propose a semantic-aware message-passing mechanism that propagates both homophilous and heterophilous messages through the edges that connect nodes with the same label. We conduct extensive experiments on 9 benchmark datasets with different heterophily ratios. Experimental results show that our method outperforms representative baselines on heterophilous datasets. Our contributions can be summarized as follows:

- We point out that existing HeterGNNs suffer from insufficient neighborhood partition and heterophily modeling, which are critical but challenging to break through.
- We propose heterophilous distribution propagation for graph neural networks (HDP), which consists of the semantic-aware neighborhood partition and heterophilous neighborhood distribution modeling to address the aforementioned challenges.
- We conduct extensive experiments to compare our models against 13 other competitors on 9 benchmark datasets with different levels of homophily. Experimental results show the superiority of our methods on the heterophilous datasets.

2. Related Work

Graph Neural Networks [18] have shown great power to model graph structured data. The representative designs [19, 20, 21, 22] aim to smooth features across the graph topology or aggregate information from neighbors and then update ego representations, both leading to the similar representations between central nodes and neighbors. However, most of them are based on an implicit assumption that the graph is homophily, while real-world graphs do not always obey it. This leads to their poor performance on heterophily graphs where nodes of different classes are connected.

Heterophilous GNNs[23, 24] have been proposed to tackle this problem. Some of them tend to **reduce the negative impact of heterophilous neighbors**. A naive idea is to decrease the proportion of heterophilous neighbors from the data level, such as neighborhood extension by high-order neighbors and graph reconstruction by node similarities. MixHop [16], a representative method, aggregates messages from multi-hop neighbors to adapt to different scales. Apart from multi-hop neighbors, UGCN [25] and SimP-GCN [26] extend the neighbor set by adding similar but disconnected nodes through the kNN algorithm. WRGAT [27] calculates the structural similarity according to the degree sequence of the neighbors and utilizes it to reconstruct a multi-relational graph. Geom-GCN [12] defines the geometric relationships to discover potential neighbors as a complement to the original neighbor set. Li et al. [28] learn discriminating node representations with the idea of spectral clustering. Based on the representation distance between nodes, a graph is reconstructed to maximize homophily. Also, some methods reduce the weights of heterophilous neighbors during aggregation from the model level. H2GCN [13] separates ego- and neighbor-representations to prevent ego-node from the pollution of heterophilous neighbors. The subsequent methods distinguish the neighbors explicitly or implicitly and set the corresponding weights. GGCN [29] distinguishes neighbors according to the signs of representation cosine similarity and applies different update weights. HOG-GCN [17] captures the pair-wise homophily estimation from attribute space and topology space and uses it as the aggregate weight. PEGFAN [30] constructs Haar-type graph framelets with the property of permutation equivariance to capture messages from multiple scales of the graph.

Further, some methods found **the advantages of heterophilous neighbors** and utilized them through high-pass filters [31, 32] or negative aggregation weights [33]. FAGCN [14] learns the attention weights of low- and high-frequency signals for each node, corresponding to the negative-available aggregate weights in the spatial domain. On this basis, ACM-GCN [15] introduces the identity filter to capture more information about the original feature. These methods capture the differences between heterophilous neighbors and central nodes and are very similar to negative sampling methods [34, 35], which dig negative samples globally. GBK-GNN [36] utilizes two kernels to capture homophilous and heterophilous information and selects the result by a gate for each node. Similarly, GloGNN [37] learns a coefficient matrix based on the self-expressiveness assumption of the linear subspace model, which guides the global message passing. UniFilter [38] proposes a polynomial filter-based graph neural network with a universal polynomial basis to tackle diverse graph heterophily degrees.

Table 1: A summary of the notations used in this paper.

Notations	Explanations	Notations	Explanations
$\mathcal{V}$	Node set	$\mathcal{E}$	Edge set
$\mathbf{X}$	Node features	$\mathbf{A}$	Adjacency matrix
$\mathbf{Y}$	Node labels	$\widehat{\mathbf{Y}}$	Predicted labels
$\mathbf{H}$	Node representations	$\mathbf{Z}$	Soft assignment
$\mathbf{A}^{hm}$	Homophilous neighborhood	$\mathbf{A}^{ht}$	Heterophilous neighborhood
$\mathbf{H}^{ego}$	Ego representation	$\mathbf{H}^{nb}$	Heterophilous neighbor distributions
$\mathbf{H}^{hm}$	Homophilous representations	$\mathbf{H}^{ht}$	Heterophilous representations
$\mathbf{X}^{nb}$	Neighbor features	$\mathbf{X}^{str}$	Structural embedding
$\mathcal{E}'$	Partitioned neighborhood	$\mathbf{P}$	Homophilous edge probabilities
$\mathcal{S}^{tra}$	Training set	$\mathcal{S}^{tgt}$	Target set
$\mathcal{S}^{tst}$	Trust set	$\mathbf{c}$	Class prototypes
N	Number of nodes	K	Number of classes
F	Feature dimension	D	Representation dimension
h	Homophily ratio	h', $\widehat{h}$	Estimated homophily ratio
λ	Rescaling parameter	ϵ	Partition threshold
κ	Random walk hop	δ	Trust set threshold
ρ	Training and validation accuracy	τ	temperature parameter
α	Attention weights	β	weight parameter

3. Preliminaries

In this section, we first give the notations and problem description, then introduce the concepts used in this paper.

3.1. Notations

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with nodes $\mathcal{V}$ and edges $\mathcal{E}$. N denotes the number of nodes. $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the adjacency matrix and $\mathbf{X} \in \mathbb{R}^{N\times F}$ is the node feature matrix with feature dimension F. Node labels are represented as $\mathbf{Y} \in \mathbb{R}^{N\times 1}$, and only a part of it is available. We use K as the number of classes and D as the representation dimension. $\mathcal{S}^{tra}$, $\mathcal{S}^{tgt}$ and $\mathcal{S}^{tst}$ denote the training set, the target set for structure encoding and trust set for trusted prototype contrastive loss respectively, which are described in the following section. Table 1 lists a summary of the used notations in this paper.

3.2. Problem Description

In this paper, we mainly focus on graph representation learning, of which the performance is evaluated by the semi-supervised node classification task. Specifically, in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each

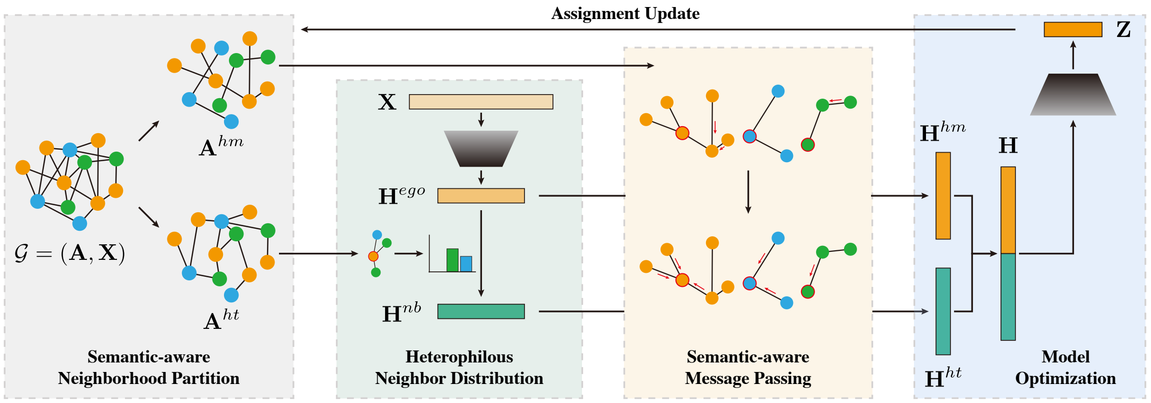


Figure 2: Overall framework of HDP, which contains three main parts including semantic-aware neighborhood partition, heterophilous neighbor distribution modeling and semantic-aware message passing.

node belongs to one of K classes and a part of labels $\mathbf{Y}$ are already known. The objective is to predict the labels of other nodes.

3.3. Homophily and Heterophily

Homophily and Heterophily are two opposite concepts related to edges, features and labels of a graph. We use the edge homophily ratio [13] $h = \frac{|(u,v)|(u,v)\in\mathcal{E} \wedge y_u = y_v|}{|\mathcal{E}|} \in [0, 1]$, the proportion of edges connecting nodes of the same class, to measure the specific homophily level for a graph. Graphs with strong homophily tend to have a high h close to 1, while graphs with strong heterophily are the opposite, i.e. $h \to 0$.

4. Methodology

In this section, we introduce our proposed HDP model in detail. An overview of HDP is given in the Figure 2. HDP first estimates the heterophily level of a graph and partitions the neighborhood as homophilous and heterophilous ones according to the semantic assignments, which are constantly refined during the training process. Then, the heterophilous neighbor distribution is modeled for each node based on the heterophilous neighborhood and an orthogonality-oriented constraint. Further, HDP propagates messages via a semantic-aware message passing mechanism to capture class-level information approximatively.

4.1. Semantic-Aware Neighborhood Partition

Most of the existing methods distinguished the neighborhood by the representation similarity and a manually selected threshold [29, 39, 40]. However, this approach lacks explainability since the definition of homophily is based on semantic labels instead of representation. Hence, we propose a semantic-aware neighborhood partition mechanism based on the soft assignments, which is a special form of labels. Specifically, we first estimate the level of heterophily in the graph as the guidance and then partition the neighborhood.

4.1.1. Heterophily Estimation

We use the homophily ratio to measure how heterophily the graph is that a lower homophily ratio means stronger heterophily. To estimate the homophily ratio of a graph, we start with an assumption that the training set shares a similar edge distribution with the full graph, which means the homophily ratio won't be too far from the truth if there are enough nodes in the training set. Therefore, the homophily ratio of the full graph can be estimated from the training set where a part of labels is available:

$$h' = \frac{|\{(u,v)|(u,v) \in \mathcal{E}' \ \wedge \ y_u = y_v \ \wedge \ u,v \in \mathcal{S}^{tra}\}|}{|\{(u,v)|(u,v) \in \mathcal{E}' \ \wedge \ u,v \in \mathcal{S}^{tra}\}|}, \tag{1}$$

where $\mathcal{E}' \in \{\mathcal{E}, \mathcal{E}^2\}$ is the partitioned neighborhood, which can be 1-hop or 2-hop considering efficiency and the number of isolated nodes in results, $\mathcal{S}^{tra}$ denotes the training set. As the estimated homophily ratio h' can't be completely accurate, we slightly rescale it to find a suitable boundary for neighborhood partition:

$$\widehat{h} = \lambda h', \tag{2}$$

where $\lambda \in [0.8, 1.2]$ is a parameter that controls the direction and strength of rescaling.

4.1.2. Neighborhood Partition

We start with the soft semantic assignments $\mathbf{Z} \in \mathbb{R}^{N \times K}$, where the sum of each row is 1 and element $\mathbf{z}_{ij}$ in row i and column j indicates the probability of node i belonging to class j. In the semi-supervised setting, the semantic assignment is obtained from the predicted results of models (e.g. the initializing MLP or whole HDP). In other words, we first train the model with training set labels and predict the soft pseudo labels for all nodes as the semantic assignments. The implementation details of semantic assignments are described in Sec 4.2.3 and 4.3.1. With the help of semantic assignments, we can construct the relationships between central nodes and neighbors in an interpretable way. Specifically, the homophilous edge probabilities $\mathbf{P}$, which indicates the probabilities that two nodes belong to the same class, can be calculated as follows:

$$\mathbf{P}_{uv} = \begin{cases} \mathbf{z}_u \mathbf{z}_v^T, & (u, v) \in \mathcal{E}', \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The neighborhood is then partitioned to fit the estimated heterophily level according to $\mathbf{P}$, i.e. a lower homophily ratio corresponds to fewer homophilous edges and vice versa. Specifically, HDP automatically generate a threshold ϵ based on $\widehat{h}$:

$$\epsilon = \mathrm{TopK}(\mathbf{P}, \widehat{h}|\mathcal{E}'|), \tag{4}$$

where $\mathrm{TopK}(\mathbf{x}, y)$ means the y-largest element in $\mathbf{x}$. Then the homophilous neighborhood $\mathbf{A}^{hm}$ and heterophilous neighborhood $\mathbf{A}^{ht}$ can be partitioned by threshold filtering:

$$\mathbf{A}^{hm}_{uv} = \begin{cases} 1, & \mathbf{P}_{uv} \geq \epsilon \ \wedge \ (u, v) \in \mathcal{E}', \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

$$\mathbf{A}^{ht}_{uv} = \begin{cases} 1, & \mathbf{P}_{uv} < \epsilon \ \wedge \ (u, v) \in \mathcal{E}', \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

To avoid the defects of outdated partition results, we conduct an **update strategy** that refines the partition results dynamically when the assignments become more accurate during the training process.

4.2. Heterophilous Neighborhood Modeling

The heterophilous neighborhood deserves to be modeled separately rather than integrated to the central nodes, since its diverse preference is crucial for the discriminability of representations. Thus, we model the heterophilous neighborhood in three steps: (1) constructing ego representation for each node, (2) modeling the heterophilous neighbor distribution of each node, and (3) propagating them via a semantic-aware message-passing mechanism.

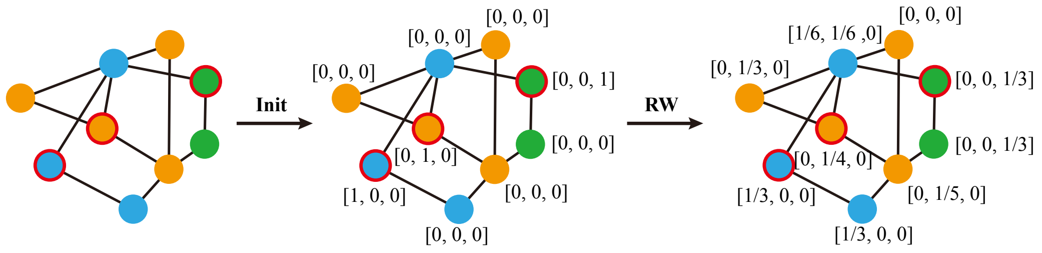


Figure 3: The illustration of semantic structural encoding. Nodes with a red circle denote the target nodes. After initialization, the structural embedding is calculated through the random walk with self-loop.

4.2.1. Ego Representation Construction

Firstly, we construct a representation for each node which contains the attribute and structural information of node ego.

To better utilize the topological information without considering the distinguishing of edges, we introduce **Semantic Structural Encoding (SSE)** as a supplementary feature alongside node attributes. Distance Encoding [41] is used where the embedding $\mathbf{X}^{str}$ denotes the landing probabilities of random walks from the corresponding node to some target nodes. Different from the existing methods that select target nodes mainly according to topology or randomness [42, 43], we construct a target node set $\mathcal{S}^{tgt}$ with the nodes in the training set $\mathcal{S}^{tra}$ to implicitly introduce some semantic information: $\mathcal{S}^{tgt} \subseteq \mathcal{S}^{tra}$. First, the structural embedding $\widehat{\mathbf{X}}_i^{str}$ is initialized as a unique one-hot vector for nodes in the target set and an all-zero vector otherwise.

$$\widehat{\mathbf{X}}_i^{str} = \begin{cases} [0, 0, ..., 1, ..., 0]\,, & i \in \mathcal{S}^{tgt} \\ [0, 0, ..., 0, ..., 0]\,, & i \notin \mathcal{S}^{tgt} \end{cases}, \tag{7}$$

where the position of '1' in the one-hot vector equals the ranking of i in $\mathcal{S}^{tgt}$. The final structural embedding $\mathbf{X}^{str}$ can be calculated by the transfer probability of random walks:

$$\mathbf{X}^{str} = (\mathbf{D}^{-1}\mathbf{A})^{\kappa}\widehat{\mathbf{X}}^{str}, \tag{8}$$

where κ is the hop number of random walk. Finally, the structural embedding captures the topological information about target nodes of different classes around the central node, which also brings semantic information. Figure 3 gives a toy example of semantic structural encoding.

The ego representations of nodes are obtained by a simple MLP with nodes' features and structural embeddings as input:

$$\mathbf{H}^{ego} = \mathrm{MLP}([\mathbf{X}\|\mathbf{X}^{str}]), \tag{9}$$

where $\|$ denotes the concatenation operation.

4.2.2. Heterophilous Neighbor Distribution

To model the distribution of heterophilous neighbors, it would be sufficient to design an operator (i.e. function) that takes heterophilous neighbors as input, with each distinct distribution of heterophilous neighbors corresponding to a different output o in the output space O, i.e.

$$\begin{gathered} O = \{o_i \mid o_i = \text{Operator}(\{\mathbf{h}_j^{ego} | \mathbf{A}_{ij}^{ht} = 1\}), i \in \mathbb{N}\} \\ \text{s.t. } \forall\ o_i, o_j \in O,\ o_i \neq o_j \text{ w.r.t. } i \neq j \end{gathered}. \tag{10}$$

On this basis, we find that the distribution of heterophilous neighbors can be effectively modeled by the mean operator, with the following two assumptions: (1) Assuming node representations exhibit clustering characteristics, where the average distance within a class is significantly smaller than the average distance between different classes. This implies that the representations of similar nodes are linearly correlated within a certain range of error. (2) Assuming the existence of a clustering center for each class's representations, referred to as a prototypes $\{\mathbf{c}_k | k \in K\}$. We have the following theorem with detailed proof:

Theorem 1. *Let Mean*$(\{\mathbf{h}_j^{ego} | \mathbf{A}_{ij}^{ht} = 1\})$ *be mean operator that aggregate heterophilous neighbor representations,* $\mathbf{c}_k$ *be the prototype of* k*-th class. Function Mean*$(\{\mathbf{h}_j^{ego} | \mathbf{A}_{ij}^{ht} = 1\})$ *is injective if it is satisfied that all class prototypes* $\mathbf{c}_k$ *are orthogonal to each other.*

The injectivity ensures that each element in the domain of the input (i.e. heterophilous neighbors' distribution) has a distinct and unique output in the output domain. We found that as long as the conditions of Theorem 1 are satisfied, the mean operator can be regarded as an injective function within a certain range of error, a simple yet effective approach to perform heterophilous neighborhood modeling.

To prove that heterophily neighbor patterns can be modeled by the mean operator, it suffices to demonstrate that the mapping function from the mean of $\{\mathbf{h}_j^{ego} | \mathbf{A}_{ij}^{ht} = 1\}$ to H^{nb} in terms of embedding is injective.

Lemma 1. *Injectivity is equivalent to null space equals* $\{0\}$*. Let* $T \in \mathcal{L}(V, W)$*,* $T(v) = T \cdot v = w$*. Then* T *is injective if and only if null*$(T) = \{0\}$*.*

Proof of lemma 1: Sufficiency: First, suppose T is injective. We want to prove that null $T = \{0\}$. We already know that $\{0\} \subset \text{null}(T)$. To prove the inclusion in the other direction, suppose $v \in \text{null}(T)$, then $T(v) = 0 = T(0)$. Because T is injective, the equation above implies that $v = 0$. Thus we can conclude that null $T = \{0\}$, as desired. **Necessity**: To prove the implication in another direction, now suppose $\text{null}(T) = \{0\}$. We want to prove that T is injective. To do this, suppose $u, v \in V$ and $T(u) = T(v)$. Then

$$0 = T(u) - T(v) = T(u - v). \tag{11}$$

Thus $u - v$ is in null T, which equals $\{0\}$. Hence $u - v = 0$, which implies that $u = v$. Hence T is injective, as desired.

Having the **Lemma 1** proofed, now we express the mean operator in the following form:

$$\mathbf{M}\mathbf{X} = b, \tag{12}$$

where M_{1*n} represents the mean operator, $\mathbf{X}_{n*D}$ is the matrix formed by embeddings of heterophilous neighbors, and b is the resulting new embedding. Assuming that embeddings of the same type of heterophilous neighbors are linearly dependent, we can rewrite this equation as:

$$\mathbf{M}^{'}\mathbf{X}_p \approx b, \tag{13}$$

where $\mathbf{M}^{'}_{1*K}$ is a weighted mean operator, $\mathbf{X}_p$ is a $K * D$ prototype embedding matrix, K is the number of classes. The injectivity of mean operator $\mathbf{M}$ involves considering the solution for $\mathbf{M}^{'}\mathbf{X}_p = 0$. It is clear that if it is satisfied that all $\mathbf{X}_p^k$ are orthogonal to each other, the null space of $\mathbf{M}^{'} = \{0\}$, indicating that the mean operator is approximately injective. In other words, each distinct input produces a unique output. Thus in this manner, the mean operator applied to heterophilous neighbors can generate distinguishable embeddings based on the distribution of heterophilous neighbors.

Hence, we approximatively model the heterophilous neighbor distribution by the mean operator with a **orthogonality-oriented constraint** to make the prototypes as orthogonal as possible, which is described in Sec 4.3. The heterophilous neighbor distributions are formatted as follows:

$$\mathbf{H}^{nb} = \mathbf{D}^{ht^{-1}}\mathbf{A}^{ht}\mathbf{H}^{ego}. \tag{14}$$

where $\mathbf{D}^{ht}$ is degree matrix with entries $\mathbf{D}^{ht}_{ii} = \sum_j \mathbf{A}^{ht}_{ij}$.

4.2.3. Semantic-Aware Message Passing

Due to the sparse nature of graph data, the heterophilous neighbor distribution of a single node could be chaotic. Hence, we introduce **Semantic-aware Message Passing (SMP)** based on the homophilous neighborhood, which aggregates information from different hops of homophilous neighbors with adaptive weights. Intuitively, nodes with similar semantics share similar characters of not only nodes themselves but also neighbor distributions. Thus, aggregating information from enough homophilous neighbors can approximatively model the class-level heterophilous neighborhood, which is more accurate and discriminative. Specifically, SMP is a multilayer module in which the messages are propagated only on the homophilous neighborhood. The l-th layer of SMP is expressed as follows:

$$\begin{aligned}\widetilde{\mathbf{H}}^l &= \mathbf{D}^{hm^{-1}}\mathbf{A}^{hm}\mathbf{H}^{(l-1)},\\ \alpha^l &= f_{\varphi^l}([\mathbf{H}^0\|\widetilde{\mathbf{H}}^l]),\\ \mathbf{H}^l &= \alpha^l\mathbf{H}^0 + (1-\alpha^{l)})\widetilde{\mathbf{H}}^l,\end{aligned} \tag{15}$$

where $\mathbf{D}^{hm}$ is a degree matrix with entries $\mathbf{D}^{hm}_{ii} = \sum_j \mathbf{A}^{hm}_{ij}$, $\mathbf{H}^0$ is the input of whole SMP, and the message from neighbors and ego nodes are linearly combined in which the weights $\alpha^l \in \mathbb{R}^{N\times 1}$ are set by a weight learner f_{φ^l}.

Then, we propagate the heterophilous neighbor distribution via SMP to capture the heterophilous distribution representations:

$$\mathbf{H}^{ht} = \mathrm{SMP}(\mathbf{H}^{nb}, l^{ht}). \tag{16}$$

Meanwhile, propagating the nodes' ego representations via SMP can capture the homophilous representations:

$$\mathbf{H}^{hm} = \mathrm{SMP}(\mathbf{H}^{ego}, l^{hm}). \tag{17}$$

l^{ht} and l^{hm} are the numbers of SMP layers. When the SMP layers become deeper, HDP can adaptively capture homophilous messages and heterophilous distribution from high-order homophilous neighbors for the central node. Specifically, in each layer, SMP aggregates messages from homophilous neighbors and combines with adaptive weights. When multiple layers are stacked, messages from high-order homophilous neighbors can be passed down through the layers with corresponding weights.

Now we have two kinds of representations that capture homophilous and heterophilous neighborhood information respectively. The final node representations are the concatenation of $\mathbf{H}^{hm}$

and $\mathbf{H}^{ht}$, which can be the input to other downstream tasks.

$$\mathbf{H} = [\mathbf{H}^{hm} \| \mathbf{H}^{ht}]. \tag{18}$$

For node classification, the soft assignments $\mathbf{Z}$ and predicted labels $\widehat{\mathbf{Y}}$ are given by a classifier f_ψ:

$$\mathbf{Z} = f_\psi(\mathbf{H}),\ \widehat{\mathbf{Y}} = \arg\max(\mathbf{Z}). \tag{19}$$

4.3. Model Training

We introduce two modules for model training, including assignment initialization and optimization. The former provides the assignment for the first neighborhood partition while the latter describes the optimization object of the whole model.

4.3.1. Assignment Initialization

HDP partitions the neighborhood by the semantic assignments, which need to be initialized to obtain a relatively accurate result at the beginning of training. For the assignment initialization, a naive approach is to train a classifier with only nodes' ego features. However, the neighbor features may also provide helpful information for classification. Since we don't know the homophily ratio of the graph, we separate the node's ego features $\mathbf{X}$ and the corresponding neighbor features $\mathbf{X}^{nb}$ to avoid pollution:

$$\mathbf{X}^{nb} = \widehat{\mathbf{A}}\mathbf{X}, \tag{20}$$

where $\widehat{\mathbf{A}}$ is the normalized adjacency matrix. The new node features are obtained through concatenating the ego features, neighbor features and structural embedding:

$$\mathbf{X}^{all} = [\mathbf{X} \| \mathbf{X}^{nb} \| \mathbf{X}^{str}]. \tag{21}$$

In practice, we choose some of them to construct the new node features according to the performance on the validation set. Finally, we can get the soft assignments $\mathbf{Z}$ through an MLP classifier trained by cross-entropy loss:

$$\mathbf{Z} = \mathrm{MLP}_{init}(\mathbf{X}^{all}). \tag{22}$$

To avoid errors caused by precision, we rescale the assignments to a relatively high level during the neighborhood partition.

4.3.2. Optimization

HDP contains two kinds of objectives: a commonly used cross-entropy loss for node classification and a Trusted Prototype Contrastive (TPC) loss as the orthogonality-oriented constraint for class prototypes.

The cross-entropy function can measure the gap between predicted results and the ground truth:

$$\mathcal{L}^{ce} = \mathrm{CE}(\mathbf{Z}, \mathbf{Y}), \tag{23}$$

where $\mathrm{CE}(\cdot,\cdot)$ denotes the cross-entropy function.

In Sec 4.2.2, we find that the modeling of heterophilous neighbor distribution needs to satisfy two conditions as far as possible: (i) node representations exhibit clustering characteristics; (2) class prototypes exist and are orthogonal to each other. These conditions are not always held in practice. Thus, we want to introduce additional constraints to satisfy them as much as possible. Many optional constraints can achieve effects severally, such as directly constraining representation similarities among nodes and prototypes. Further, is there a constraint that can achieve both effects at the same time? We introduce the Trusted Prototype Contrastive (TPC) loss, which is inspired by the original prototype contrastive learning (PCL) [44].

We first introduce TPC loss in detail, and then analyze how it meets the requirements. Contrastive learning [22] aims to pull positive samples together while pushing negative samples away. As a variant, PCL constructs positive and negative samples between samples and class prototypes calculated by the pseudo labels, leading to highly discriminative representations. In our TPC loss, we first select some high confidence nodes as trust set $\mathcal{S}^{tst}$:

$$\mathcal{S}^{tst} = \{v_i | \mathbf{Z}_i^{max} \geq \delta\}, \tag{24}$$

where $\mathbf{Z}^{max}$ is the maximum value in each row of $\mathbf{Z}$, denoting the largest probability of each node belonging to any class. δ is a threshold decided by the accuracy ρ of training and validation set:

$$\delta = \mathrm{TopK}(\mathbf{Z}^{max}, \rho|\mathcal{V}|). \tag{25}$$

Then, the prototype of class j can be calculated as the mean of node ego representations $\mathbf{H}^{ego}$ within trust set:

$$\begin{gathered} \mathbf{c}_j = \frac{1}{|\mathcal{S}_j^{tst}|} \sum_{v_i \in \mathcal{S}_j^{tst}} \mathbf{h}_i^{ego}, \\ \mathcal{S}_j^{tst} = \{v_i | v_i \in \mathcal{S}^{tst} \wedge \widehat{\mathbf{Y}}_i = j\}. \end{gathered} \tag{26}$$

Formally, the trusted prototype contrastive loss can be expressed as follows:

$$\mathcal{L}^{tpc} = -\sum_{v_i \in \mathbf{S}^{tru}} \log \frac{\exp(s(\mathbf{h}_i^{ego}, \mathbf{c}_j)/\tau)}{\sum_{k=1}^{K} \exp([s(\mathbf{h}_i^{ego}, \mathbf{c}_k)]_+/\tau))}, \tag{27}$$

where τ is a temperature parameter, $[\cdot]_+ = \max(\cdot, 0)$ and $s(\cdot, \cdot)$ is cosine similarity function:

$$s(\mathbf{h}_i^{ego}, \mathbf{c}_j) = \frac{\mathbf{h}_i^{ego} \cdot \mathbf{c}_j}{|\mathbf{h}_i^{ego}||\mathbf{c}_j|}. \tag{28}$$

The numerator of Eq 27 indicates the positive pairs, i.e. nodes and corresponding prototypes, while the denominator indicates the negative pairs, i.e. nodes and other prototypes. The goal of the constraint is to make the numerator as large as possible and the denominator as small as possible. Since the prototypes are computed from representations of the same class, TPC loss can pull nodes from the same class together and push nodes from different classes away, which satisfies the first condition. In the ideal case, the cosine similarity will be 1 in the numerator and 0 in the denominator since it's set to be non-negative during optimization. In other words, the optimal case of TPC loss is that all class prototypes are orthogonal to each other, indicating that the TPC loss is also an orthogonality-oriented constraint as we desired. Thus, we have the following remark:

Remark 1. *TPC loss can provide corresponding constraints for the two conditions required by heterophilous neighbor distribution modeling at the same time, which is helpful for the handling of heterophilous graphs.*

Finally, the overall optimization objective can be written as follows:

$$\mathcal{L} = \mathcal{L}^{ce} + \beta\mathcal{L}^{tpc}. \tag{29}$$

where β is a weight parameter. Finally, we summarize the whole process of HDP in Algorithm 1.

5. Experiments

In this section, we first evaluate the representation learning performance of HDP through node classification task against some state-of-the-art methods on 9 public datasets. Then, the effectiveness of components in HDP is shown by an ablation study and some visualizations.

Algorithm 1 Algorithm of HDP

```
Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, training set $\mathcal{S}^{tra}$, node labels $\mathbf{Y}$, adjacency matrix $\mathbf{A}$, node features $\mathbf{X}$, rescaling parameter $\lambda$, epoch $E$
Ensure: Predicted labels $\hat{\mathbf{Y}}$
1: Construct structural embedding $\mathbf{X}^{str}$ via Eq.7 and Eq.8.
2: Initialize the assignment $\mathbf{Z}$ via Eq.22.
3: Estimate the homophily ratio of graph via Eq.1 and Eq.2.
4: Partition the neighborhood to $\mathbf{A}^{hm}$ and $\mathbf{A}^{ht}$ via Eq.5 and Eq.6.
5: Establish trust set $\mathcal{S}^{tru}$ via Eq.24.
6: for iteration 1, 2, ..., E do
7:     Construct ego representations $\mathbf{H}^{ego}$ for nodes via Eq.9.
8:     Modeling heterophilous neighbor distribution $\mathbf{H}^{nb}$ via Eq.14.
9:     Propagate $\mathbf{H}^{nb}$ to capture the heterophilous distribution representations $\mathbf{H}^{ht}$ via Eq.16
10:    Propagate $\mathbf{H}^{ego}$ to capture the homophilous representations $\mathbf{H}^{hm}$ via Eq.17
11:    Obtain final representations $\mathbf{H}$, assignments $\mathbf{Z}$ and predicted labels $\widehat{\mathbf{Y}}$ via Eq.18 and Eq.19.
12:    Calculate loss $\mathcal{L}$ via Eq.27 and Eq.29.
13:    Back-propagation $\mathcal{L}$ to optimize the weights of networks.
14:    if current assignment $\mathbf{Z}$ performs better then
15:        update the neighborhood partition results $\mathbf{A}^{hm}$ and $\mathbf{A}^{ht}$ via Eq.5 and Eq.6 with current $\mathbf{Z}$.
16:        update the trust set $\mathcal{S}^{tst}$ via Eq.24 with current $\mathbf{Z}$.
17:    end if
18: end for
19: return $\widehat{\mathbf{Y}}$
```

Table 2: Detailed statistics of datasets and classification performance on 6 datasets with various levels of heterophily and 3 homophilous datasets.

Methods	**Cornell**	**Texas**	**Wisconsin**	**Chameleon**	**Actor**	**Squirrel**	**Cora**	**Citeseer**	**Pubmed**
Homo. Ratio	0.3	0.11	0.21	0.23	0.22	0.22	0.81	0.74	0.8
#Nodes	183	183	251	2277	7600	5201	2708	3327	19717
#Edges	280	295	466	31421	26752	198493	5278	3703	44327
#Features	1703	1703	1703	2325	931	2089	1433	3703	500
#Classes	5	5	5	5	5	5	7	6	3
MLP	84.32 ± 4.32	82.43 ± 6.76	85.10 ± 3.53	48.27 ± 1.91	36.70 ± 1.09	34.30 ± 0.99	73.76 ± 1.95	72.87 ± 2.07	87.39 ± 0.34
GCN	61.08 ± 5.43	62.16 ± 4.83	60.98 ± 5.15	67.19 ± 2.16	30.31 ± 1.19	51.71 ± 1.18	86.94 ± 0.97	75.92 ± 1.49	87.64 ± 0.54
GAT	59.46 ± 3.63	64.59 ± 5.47	60.78 ± 8.68	63.38 ± 1.25	30.37 ± 0.86	54.07 ± 1.85	85.37 ± 1.46	74.92 ± 1.70	86.38 ± 0.48
GCNII	77.86 ± 3.79	77.57 ± 3.83	80.39 ± 3.40	63.86 ± 3.04	37.44 ± 1.30	38.47 ± 1.58	**88.37 ± 1.25**	$\underline{77.33}$ ± 1.48	**90.15 ± 0.43**
MixHop	81.08 ± 6.28	82.97 ± 6.17	84.31 ± 3.16	66.18 ± 1.57	34.81 ± 0.58	56.26 ± 1.69	86.48 ± 1.04	76.72 ± 1.07	88.39 ± 0.45
H_2GCN	82.16 ± 4.80	84.86 ± 6.77	86.67 ± 4.69	59.39 ± 1.98	35.86 ± 1.03	37.90 ± 2.02	87.67 ± 1.42	77.07 ± 1.64	89.59 ± 0.33
UGCN	73.78 ± 5.41	75.95 ± 5.85	78.43 ± 4.96	63.22 ± 2.01	32.55 ± 1.36	50.09 ± 2.91	72.43 ± 2.75	74.53 ± 1.38	80.73 ± 1.36
WRGAT	81.62 ± 3.90	83.62 ± 5.50	86.98 ± 3.78	65.24 ± 0.87	36.53 ± 0.77	48.85 ± 0.78	88.20 ± 2.26	76.81 ± 1.89	88.52 ± 0.92
GPR-GNN	79.46 ± 6.30	83.51 ± 5.98	83.53 ± 4.04	69.08 ± 2.41	35.22 ± 1.00	52.06 ± 1.31	87.77 ± 1.14	75.98 ± 1.53	86.98 ± 0.49
LINKX	77.84 ± 5.81	74.60 ± 8.37	75.49 ± 5.72	68.42 ± 1.38	36.10 ± 1.55	$\underline{61.81 \pm 1.80}$	84.64 ± 1.13	73.19 ± 0.99	87.86 ± 0.77
GGCN	85.68 ± 6.63	84.86 ± 4.55	86.86 ± 3.29	71.14 ± 1.84	$\underline{37.54 \pm 1.56}$	55.17 ± 1.58	87.95 ± 1.05	77.14 ± 1.45	89.15 ± 0.37
ACM-GCN	85.14 ± 6.07	$\underline{87.84 \pm 4.40}$	$\underline{88.43 \pm 3.22}$	69.14 ± 1.91	36.63 ± 0.84	55.19 ± 1.49	87.91 ± 0.95	77.32 ± 1.70	$\underline{90.00 \pm 0.52}$
GloGNN	$\underline{85.95 \pm 5.10}$	84.32 ± 4.15	88.04 ± 3.22	$\underline{71.21 \pm 1.84}$	**37.70 ± 1.40**	57.88 ± 1.76	$\underline{88.33 \pm 1.09}$	**77.41 ± 1.65**	89.62 ± 0.35
HDP	**87.84 ± 4.23**	**88.38 ± 4.20**	**88.82 ± 3.40**	**71.56 ± 2.52**	37.26 ± 0.67	**62.07 ± 1.57**	87.00 ± 1.35	77.10 ± 1.56	89.49 ± 0.47

5.1. Datasets and Baselines

Experiments are conducted on 6 public heterophilous graph datasets including Cornell, Texas, Wisconsin, Chameleon, Actor, and Squirrel [12], and 3 homophilous datasets including Cora, Citeseer and Pubmed [45]. The detailed statistics of these datasets are summarized in Table 2 while the descriptions are in follows:

- Cornell, Texas and Wisconsin are three sub-datasets of the webpage dataset collected from computer science departments of various universities, where nodes are web pages belonging to one of five categories, and edges represent the hyperlinks between them.
- Chameleon and Squirrel are two webpage networks in Wikipedia. The nodes are classified into five categories based on their average amounts of monthly traffic.
- Actor (also named Film) is a subgraph of the film-director-actor-writer network, where nodes are actors and edges denote the co-occurrence relation between them in Wikipedia pages.
- Cora, Citeseer and Pubmed are citation networks with high homophily. In these datasets, nodes represent the scientific papers while edges denote citations. The node label is the research field of a paper.

We compare HDP with 13 baseline methods, including (1) MLP; (2) general GNN methods: GCN [19], GAT [21] and GCNII [22]; (3) methods adapted to heterophilous graphs: MixHop [16], H_2GCN [13], UGCN [25], WRGAT [27], GPR-GNN [46], LINKX [47], GGCN [29], ACM-GCN [15] and GloGNN [37]. The first six heterophilous GNN methods tend to reduce the negative impact of heterophilous neighbors while the last three utilize the difference between ego node and heterophilous neighbors.

5.2. Experimental Settings

We implement HDP by PyTorch and run experiments on the Nvidia RTX 3090 GPU. The models are optimized by Adam [48]. For the hyperparameter setting, we use an anneal strategy to turn the hyperparameter combination based on the results of the validation set. Early stop strategy is applied for model training with the parameter ”patience”. The assignment initialization is separated from the main part of HDP for parameter tuning. Detailed search space of hyperparameters is listed in Table 3, while specific settings can be seen in codes.

Table 3: The search space and settings of hyper-parameters.

Notation	Range
learning_rate_init	{0.001, 0.003, 0.01, 0.03}
weight_decay_init	{0, 1e-5, 5e-5, 1e-4, 5e-4, 0.001, 0.005}
epoch_init	{500, 1000}
patience_init	{50, 100, 200, 400}
structural_dim	{0, 2^6, 2^7, ..., 2^{13}}
hidden_dim	{512}
embedding_dim	{128}
learning_rate	{0.0003, 0.001, 0.003, 0.01, 0.03}
weight_decay	{0, 5e-6, 5e-5, 5e-4, 0.001, 0.005, 0.01}
epoch	{2000}
patience	{50, 100, 200, 400}
order	{1, 2}
β	{0.1, 1, 10}
τ	{0.1, 0.2, 0.5}
λ	[0.8, 1.2]
κ	[0, 8]
l^{hm}	[0, 8]
l^{ht}	[0, 8]

For a fair comparison, we run the experiments with the same split (48%/32%/20% of nodes for train/validation/test) of datasets from previous papers [12, 37], and report the average accuracy and corresponding standard deviation score over 10 runs on different splits. Since the results of some baseline methods on these datasets are public, we directly report them. For methods with absent results on some datasets, we use the official code released by corresponding authors and finetune the parameters as suggested in the original paper.

5.3. Performance

Table 2 shows the semi-supervised node classification performance results of all the methods on 9 benchmark datasets. We highlight the top-rank and rank two results among all methods in bold and underlining respectively for all datasets. From the table, we have the following observations:

MLP shows the basic baseline performance, which only uses the feature of node ego for classification. It performs well on most heterophilous datasets especially on Actor, indicating the important role of ego feature for node classification on both homophilous and heterophilous datasets.

General GNN methods aggregate the neighbors' information to the central node, which brings performance improvement on homophilous datasets. However, this also leads to the pool performance on heterophilous datasets since a large number of heterophilous neighbors can contaminate the node representation. As for the improvement on Chameleon and Squirrel compared with MLP, we believe that the neighbors' information contributes much more than the central node to classification in these two datasets. Thus the neighbors' information, no matter homophilous or heterophilous, can achieve higher performance. Compared with GCN and GAT, GCNII generally performs better since the initial residual and identity mapping mechanisms implicitly combine intermediate representations and reduce the influence of neighbors, which are fit for heterophilous graphs.

Predictably, the **heterophilous GNN** methods perform relatively well on the heterophilous datasets. As early methods, MixHop, H_2GCN, UGCN, GPR-GNN and WRGAT seek for higher homophily while reducing the negative impact of heterophilous neighbors. Thus, they also show a slight improvement in homophilous datasets compared with primary HomoGNNs. However, there are still quite a number of heterophilous neighbors which keep them away from the best performance. LINKX achieves excellent performance on Squirrel but works badly on others. This is probably because the mechanism of separating then mixing adjacency and feature information is more applicable to Squirrel. Further, GGCN, ACM-GCN and GloGNN utilize the difference between the central node and heterophilous neighbors, which brings additional information for classification. As a result, they achieve better performance in both heterophilous and homophilous datasets.

HDP achieves the best results in most heterophilous datasets except Actor, demonstrating the effectiveness of semantic-aware neighborhood partition and the heterophilous neighbor distribution modeling. For Actor, we believe the unsatisfactory result is due to the inaccurate neighborhood partition which is limited by the classification performance. For homophilous datasets Cora, Citeseer and Pubmed, the helpful information from heterophilous neighbors is relatively little, which is further shown in Sec 5.9. The detailed reason is analyzed in Sec 5.10. Thus, the performance of HDP on homophilous datasets is not the best but also reaches the first echelon.

5.4. Ablation Study

HDP contains some important components that may have a significant impact on the classification performance. To show the contribution of each component to the model, we conduct

Table 4: Ablation study of HDP's main components on representative datasets.

Methods	**Wisconsin**	**Actor**	**Squirrel**	**Citeseer**
Init.	86.86 ± 3.29	37.24 ± 0.95	61.62 ± 1.71	75.94 ± 1.21
HDP w/o Init.	62.75 ± 13.15	26.94 ± 3.14	31.60 ± 3.24	72.87 ± 2.22
HDP w/o Homo.	67.45 ± 5.13	35.55 ± 0.82	61.05 ± 1.50	46.02 ± 5.88
HDP w/o Hete.	86.47 ± 4.25	36.79 ± 0.61	61.60 ± 1.81	77.07 ± 1.51
HDP w/o SSE.	85.10 ± 2.51	37.26 ± 0.67	59.75 ± 2.24	76.58 ± 1.47
HDP w/o SMP.	86.86 ± 4.96	36.44 ± 0.91	55.98 ± 1.75	75.00 ± 1.78
HDP w/o TPC.	74.90 ± 16.28	36.36 ± 1.03	55.55 ± 12.69	73.76 ± 1.53
HDP w/o Upd.	87.84 ± 2.60	37.18 ± 0.73	61.97 ± 1.59	76.96 ± 1.43
HDP	**88.82 ± 3.40**	**37.26 ± 0.67**	**62.07 ± 1.57**	**77.10 ± 1.56**

an ablation study on four representative datasets. Specifically, we explore the role of the assignment initialization module (Init), semantic structural encoding (SSE), homophilous representations (Homo), heterophilous distribution representations (Hete), semantic-aware message passing mechanism (SMP), trusted prototype contrastive loss (TPC) and the update strategy (Upd). "Init." denotes the results of the assignment initialization module. For HDP without assignment initialization, we use a randomly initialized MLP without training to construct initial assignments. Meanwhile, for HDP without homo-/hetero-philous representations, we only pass the other one as the input of the classifier. The results are shown in Table 4. From the overall level, all seven components have a positive contribution to the model. Specifically, we have the following observations and analysis:

- The **assignment initialization module** achieves satisfactory results as a separate module and provides a good foundation for HDP. Without initialization, HDP has some obvious performance reduction since the initial partition could be extremely inaccurate, which leads to error accumulation. Fortunately, the update strategy can gradually fix some errors and thus avoid the model collapse.

- Relatively speaking, the **homophilous representations** provide more performance gain than the heterophilous distribution representations, which fits the intuition. Meanwhile, the

heterophilous neighborhood representations are also effective on heterophilous datasets especially Squirrel since the absence of homophilous representations only brings slight performance reduction.

- The **semantic structural encoding** provides additional topology and semantic information for representation learning thus improving the performance. Distinctively, it doesn't seem to be helpful to Actor, since the 0-dimension structural embedding performs best.
- The **semantic-aware message passing** propagates representations along homophilous edges and overcomes the limited neighborhood distribution of a single node caused by the sparse nature of graphs. This produces class-unified and more discriminative representations.
- The **trusted prototype contrastive loss** plays a key role in the overall model since it brings discriminability for both homophilous representations and heterophilous distribution representations via the orthogonality-oriented constraint. Without TPC loss, the premise of heterophilous distribution modeling won't be hold, which leads to significant performance degradation.
- During the training process, the **update strategy** creates a virtuous cycle between neighborhood partitions and assignment accuracy. The performance improvement also shows the effectiveness of the mutually enhanced optimization between representation learning and neighborhood partition.

5.5. Complexity Study

In this section, we first analyze the space and time complexity of HDP and then conduct experiments on a large graph to discuss the practicability of HDP in real scenarios.

Regarding space requirements, the number of learnable parameters in HDP contains three main groups: initializing MLP, backbone and SMP, with up to $(2F + |\mathcal{S}^{tgt}| + K)D$, $(2F + |\mathcal{S}^{tgt}|)D$ and $2(l^{hm} + l^{ht})D$ parameters respectively. Thus, HDP contains $(4F + 2|\mathcal{S}^{tgt}| + K + 2(l^{hm} + l^{ht}))D$ parameters. For brevity, the space complexity is $O((F + |\mathcal{S}^{tgt}| + K + l^{hm} + l^{ht})D)$.

The time complexity of HDP (without initialization) is composed of three parts: (i) Neighborhood partition: $O(M)$ for homophily ratio estimation, $O(KM)$ for homophilous edge probability

Table 5: The classification performance and training time cost of representative methods on Ogbn-Arxiv.

Methods	**MLP**	**GCN**	**GAT**	**GCNII**	**HDP**
Performance	42.94 ± 1.71	56.71 ± 3.34	55.30 ± 7.05	**68.05 ± 0.20**	63.90 ± 0.65
Training time (s)	549	1193	3109	61076	7158

calculation and $O(M)$ for partition, where $M = |\mathcal{E}|$ denotes the number of edges; (ii) Heterophilous neighbor distribution construction: $O((F+|\mathcal{S}^{tgt}|)ND)$ for ego representation and $O(MD$ for neighbor distribution construction; (iii) Semantic-aware message passing: $O((N+M)DL)$ for L-layer SMP. As a result, the overall time complexity is $O(KM + (F+|\mathcal{S}^{tgt}|)ND + (N+M)DL)$.

Since the sizes of graphs are large in real scenarios, we choose a large graph, Ogbn-Arxiv with 169,343 nodes and 1,166,243 edges, as an example to conduct experiments. While dealing with large graphs, HDP focuses on edges for computation rather than nodes to limit memory requirements, which utilizes the sparsity of graphs. Specifically, the pair-wise similarities are only calculated within connected nodes, and the message passing is based on the sparse matrix. Table 5 shows the results of HDP compared with general baselines on Ogbn-Arxiv. Although GCNII can achieve the best results, it consumes too much time and thus is not suitable for real scenarios. As a comparison, HDP strikes a balance between effectiveness and time consumption, achieving good results while maintaining acceptable time consumption given the cost of GCN and GAT. This provides proof for the practicability of HDP in real scenarios.

5.6. Impact of Various Numbers of Labels for Training

The number of training labels greatly impacts the accuracy of pseudo labels, which further affects the quality of semantic-aware neighborhood partition and node representation learning. However, in many practical application scenarios, only a small proportion of node labels are available since the cost of obtaining labels is very high. Thus, we conduct an experiment on the impact of various ratios of available node labels. Specifically, we add four additional training ratios (5%, 10%, 20%, 30%) with 10 random splits each, the same set as the main ratio (48%). The test ratio is fixed as 20% while other nodes are regarded as the validation set.

Table 6: Classification performance of HDP with various numbers of training labels on representative datasets.

Training Ratio	**Wisconsin**	**Actor**	**Squirrel**	**Citeseer**
#Nodes	251	7600	5201	3327
5%	77.55 ± 2.74	34.33 ± 1.52	36.91 ± 2.11	72.02 ± 1.47
10%	84.08 ± 2.00	35.21 ± 1.32	43.17 ± 1.83	72.91 ± 2.16
20%	84.29 ± 2.59	36.47 ± 1.61	50.08 ± 1.17	74.45 ± 1.15
30%	86.94 ± 3.44	36.68 ± 0.98	55.55 ± 1.35	75.49 ± 0.90
48%	**88.82 ± 3.40**	**37.26 ± 0.67**	**62.07 ± 1.57**	**77.10 ± 1.56**

Table 6 shows the average classification performances on various training ratios. As expected, all performance improved when the training ratios increased. For Wisconsin, Actor and Citeseer, HDP can perform well with only 5% training labels. Noticed that in Wisconsin, 5% training labels means that only about 2 nodes have labels for each class. In this situation, HDP can also achieve acceptable performance and significantly outperform GCN. For Squirrel, the training ratio of labels has a relatively great influence on the performance. This may be because homophilous messages and heterophilous distributions in Squirrel, although discriminative, can easily interfere with each other. Thus, initial low-quality neighborhood distribution can lead to confusing node representations, which reduces the effectiveness of iterative enhancement between neighborhood partition and representation learning.

5.7. The Results of Neighborhood Partition

We show the results of neighborhood partition in Figure 4. The first column of each dataset shows the ratio of homophilous edges and heterophilous edges in the whole neighborhood, where the dividing line corresponds to the homophily ratio of the graph. Note that the original graphs are processed by some operations such as undirected graph conversion and adding self-loop. Thus the homophily ratio of the processed graph may be different from the original graph. The second and third columns denote the partitioned homophilous and heterophilous neighborhoods respectively.

Although the accuracy of neighborhood partition is limited by incomplete labels, it still shows great power to handle heterophily thanks to the pseudo assignments. For Wisconsin, the estimated homophily ratio is almost accurate and the partition result is also quite correct thanks to

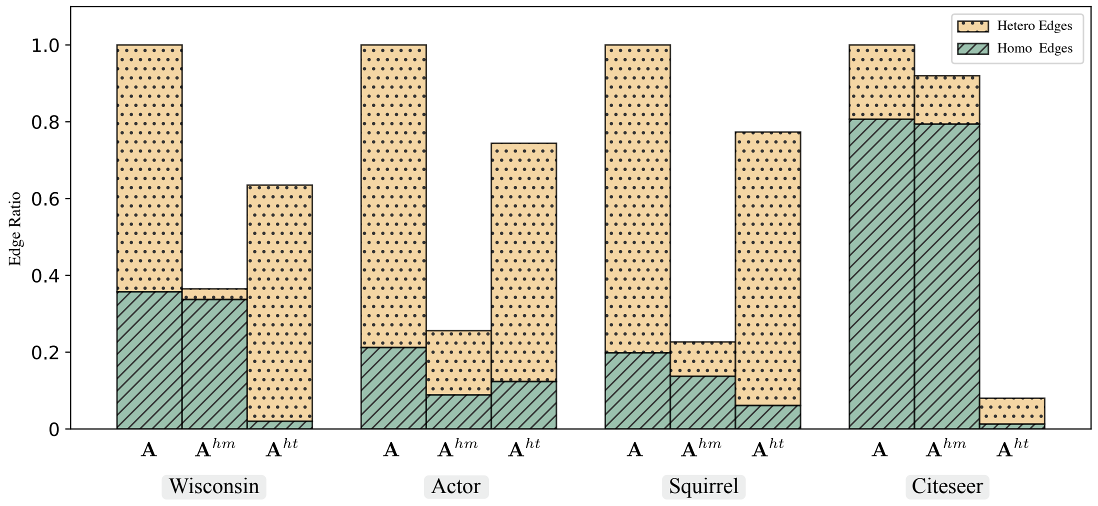
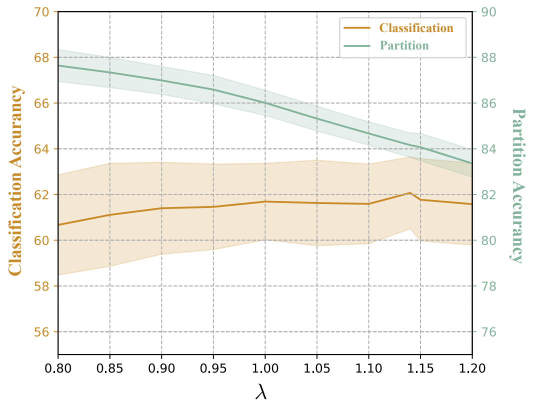


Figure 4: The visualization of neighborhood partition results on representative datasets. The 3 columns of each dataset denote the whole neighborhood, the homophilous and heterophilous neighborhoods partitioned by HDP respectively.

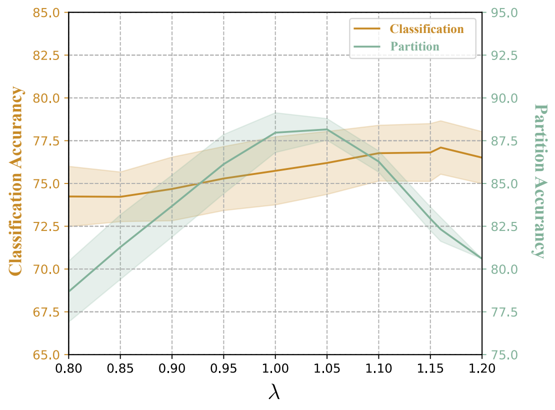
the high accuracy of assignments. Actor and Squirrel have similar heterophily levels, but their partition results are quite different because of the classification accuracy. For Actor, although the heterophily of the homophilous neighborhood has been reduced, it still suffers from the limitation of unsatisfactory accuracy. As a result, a large number of homophilous neighbors are partitioned as heterophilous, which leads to inaccurate distribution modeling and further affects the classification performance as we analyzed in Sec 5.3. For Squirrel, some homophilous edges are incorrectly partitioned, but the homophilous neighborhood is getting better since the heterophily level is reduced. For Citeseer, there is a gap between the estimated homophily ratio and the truth. But it is also effective since the homophily ratio becomes higher in the homophilous neighborhood and very low in the heterophilous neighborhood. To sum up, the adaptive neighborhood partition mechanism can adapt to different levels of heterophily and produce high-quality neighborhoods.

5.8. Influence of Rescaling Parameter

The rescaling parameter λ controls the trade-off between accuracy and recall of neighborhood partition. A low λ makes the model choose high-confidence edges as the homophilous neighborhood while abandoning edges that could be homophilous but with relatively low confidence, and vice versa. To show the impact of λ, we give the changes of partition accuracy and node classification



(a) Squirrel



(b) Citeseer

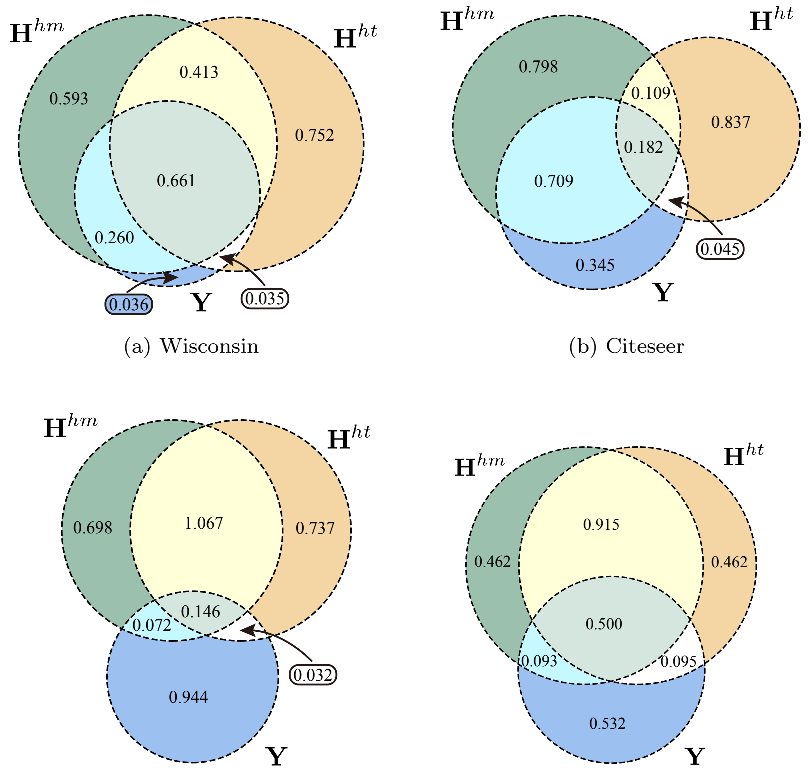
Figure 5: The variation of partition and classification accuracy while changing rescaling parameter λ. We calculate the average and standard deviation of 10 split/runs which are shown by the lines and shaded area.

accuracy concerning λ in Fig 5. Specifically, partition accuracy is evaluated by regarding the neighborhood partition as a binary classification problem.

As we expected, the variation of the neighborhood partition shows an inverted U-shaped curve as in Figure 5b. This also illustrates that the rescaling parameter λ is important since estimating graph homophily from the training set may be inaccurate as we said before. For Squirrel, λ needs to be smaller to show the other half of the curve. As for the classification accuracy, it's quite steady as λ changes, which shows the robustness of HDP. Further, the best point of the two kinds of accuracy is not the same. A large λ seems to be better for the classification accuracy. This can be attributed to the semantic-aware message passing, for which a relatively complete homophilous neighborhood is desired.

5.9. Contribution of Homophilous and Heterophilous Representations

To intuitively observe the contribution of heterophilous distribution, we estimate the mutual information between two kinds of representations and labels via DIM [49]. The results are shown in Figure 6 as a Venn diagram, where the overlap between two circles denotes the value of corresponding mutual information, and big mutual information value means an important role. We have the first interesting observation that the overlap between the label circle and others corresponds to



(a) Wisconsin (b) Citeseer

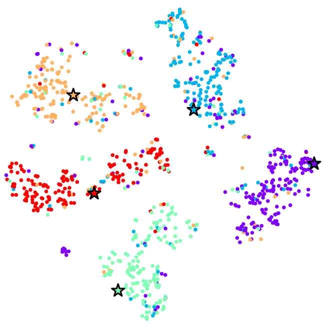
(c) Actor (d) Squirrel

Figure 6: Mutual information between two kinds of Representations and Labels

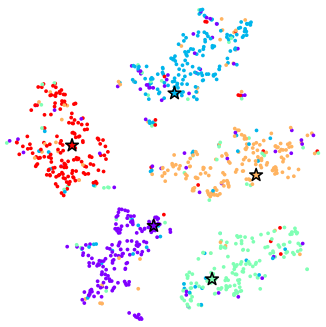
the classification performance. The bigger the overlap area is, the bigger the mutual information between labels and representations is, thus the better the classification performance is. For Wisconsin and Citeseer, homophilous representations play a more important role in node discrimination. This is consistent with intuition, especially in homophilous datasets like Citeseer. For Actor and Squirrel, the heterophilous representations show similar contributions with homophilous ones. It illustrates that our heterophily modeling is helpful in handling heterophilous graphs.

5.10. Visualization of Representation Discriminability

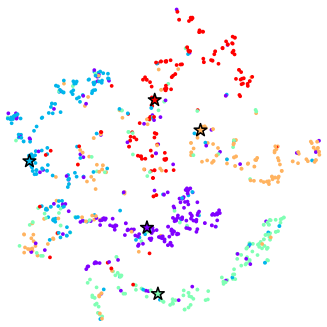
To prove the effectiveness of trusted prototype contrastive loss on the prototype orthogonality, which brings the discriminability to representations, we visualize the node representations and prototypes via T-SNE [50] in Figure 7. The classes of nodes are shown in different colors while the



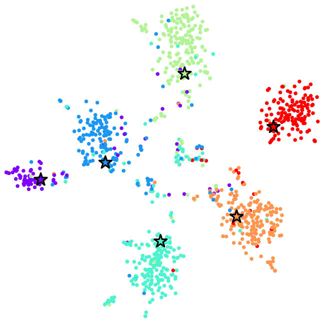
(a) $\mathbf{H}$ of Squirrel



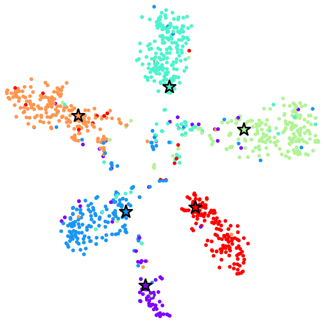
(b) $\mathbf{H}^{hm}$ of Squirrel



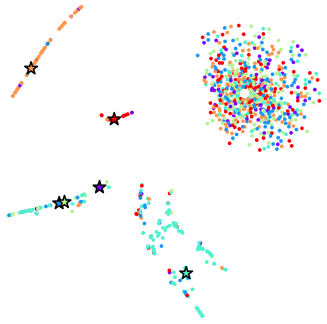
(c) $\mathbf{H}^{ht}$ of Squirrel



(d) $\mathbf{H}$ of Citeseer



(e) $\mathbf{H}^{hm}$ of Citeseer



(f) $\mathbf{H}^{ht}$ of Citeseer

Figure 7: The T-SNE visualization of representations and prototypes.

stars denote the prototypes. The results of final representations $\mathbf{H}$, homophilous representations $\mathbf{H}^{hm}$ and heterophilous distribution representations $\mathbf{H}^{ht}$ show the effectiveness of HDP, TPC loss and heterophily modeling respectively.

From Figure 7a, 7b, 7d and 7e, we can see clear boundaries between classes, which indicates the high discriminability of representations. Further, they also signify TPC loss well constrains the orthogonality of representations and HDP learns high-quality representations. In Figure 7c, the result also shows discrimination. Note that the heterophilous distribution representations $\mathbf{H}^{ht}$ are constructed without the node's ego feature. Thus, the discrimination of $\mathbf{H}^{ht}$ illustrates our heterophily modeling is effective and captures additional discriminative information from heterophilous neighbors as we desire. In Figure 7f, the result looks like a bit of a mess. Some representations have clear boundaries between classes while others mix. This may be due to the quantity of heterophilous neighbors of nodes in the homophilous dataset Citeseer being too small. Although we solve this problem with the semantic-aware message passing, there still are some classes that don't have clear connection preferences to other classes. This is a limitation of HDP: if there is no clear connection preference between classes, the heterophily modeling is unable to capture discriminative representations.

6. Discussion and Conclusion

In this paper, we study the problem of Heterophlous Graph Neural Networks (HeterGNNs), which is important in real-world graph mining scenarios. To overcome the shortcomings of existing methods in insufficient neighborhood partition and heterophily modeling, we propose Heterophilous Distribution Propagation for Graph Neural Networks (HDP). Specifically, HDP adaptively separates the neighbors into homophilous and heterphilous parts based on the pseudo assignments during training and propagates both homophilous patterns and heterophilous neighborhood distribution with a novel semantic-aware message passing module. Extensive experiments on 9 real-world datasets demonstrate the effectiveness of the HDP method. On the other hand, HDP has a limitation that the nodes should have connection preferences to nodes from other classes. Otherwise, the heterophilous distribution will lose its discriminability. In our future works, we will explore more advanced distribution modeling and more efficient model updating strategies.

7. Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant No.62106221, 62372408), Zhejiang Provincial Natural Science Foundation of China (Grant No. LTGG23F030005), and Ningbo Natural Science Foundation (Grant No.2022J183).

References

[1] S. Xiao, S. Wang, Y. Dai, W. Guo, Graph neural networks in node classification: survey and evaluation, Machine Vision and Applications 33 (2022) 1–19.

[2] B. Tang, X. Chen, S. Wang, Y. Xuan, Z. Zhao, Generalized heterophily graph data augmentation for node classification, Neural Networks 168 (2023) 339–349. `doi:https://doi.org/10.`

1016/j.neunet.2023.09.021.
URL https://www.sciencedirect.com/science/article/pii/S0893608023005142

[3] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A: statistical mechanics and its applications 390 (6) (2011) 1150–1170.

[4] G. Salha-Galvan, J. F. Lutzeyer, G. Dasoulas, R. Hennequin, M. Vazirgiannis, Modularity-aware graph autoencoders for joint community detection and link prediction, Neural Networks 153 (2022) 474–495. doi:https://doi.org/10.1016/j.neunet.2022.06.021.
URL https://www.sciencedirect.com/science/article/pii/S0893608022002362

[5] H. Cai, V. W. Zheng, K. C.-C. Chang, A comprehensive survey of graph embedding: Problems, techniques, and applications, IEEE transactions on knowledge and data engineering 30 (9) (2018) 1616–1637.

[6] W. Ju, X. Luo, Z. Ma, J. Yang, M. Deng, M. Zhang, Ghnn: Graph harmonic neural networks for semi-supervised graph-level classification, Neural Networks 151 (2022) 70–79. doi:https://doi.org/10.1016/j.neunet.2022.03.018.
URL https://www.sciencedirect.com/science/article/pii/S089360802200096X

[7] X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, L. Akoglu, A comprehensive survey on graph anomaly detection with deep learning, IEEE Transactions on Knowledge and Data Engineering.

[8] D. Guo, Z. Liu, R. Li, Regraphgan: A graph generative adversarial network model for dynamic network anomaly detection, Neural Networks 166 (2023) 273–285. doi:https://doi.org/10.1016/j.neunet.2023.07.026.
URL https://www.sciencedirect.com/science/article/pii/S0893608023003842

[9] A. F. Al Musawi, S. Roy, P. Ghosh, Identifying accurate link predictors based on assortativity of complex networks, Scientific Reports 12 (1) (2022) 18107.

[10] S. Feng, H. Wan, N. Wang, M. Luo, Botrgcn: Twitter bot detection with relational graph convolutional networks, in: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2021, pp. 236–239.

[11] Y. Shen, X. Jiang, Z. Li, Y. Wang, C. Xu, H. Shen, X. Cheng, Uniskgrep: A unified representation learning framework of social network and knowledge graph, Neural Networks 158 (2023) 142–153. `doi:https://doi.org/10.1016/j.neunet.2022.11.010`.
URL `https://www.sciencedirect.com/science/article/pii/S0893608022004518`

[12] H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, B. Yang, Geom-gcn: Geometric graph convolutional networks, arXiv preprint arXiv:2002.05287.

[13] J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, D. Koutra, Beyond homophily in graph neural networks: Current limitations and effective designs, Advances in Neural Information Processing Systems 33 (2020) 7793–7804.

[14] D. Bo, X. Wang, C. Shi, H. Shen, Beyond low-frequency information in graph convolutional networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 3950–3957.

[15] S. Luan, C. Hua, Q. Lu, J. Zhu, M. Zhao, S. Zhang, X.-W. Chang, D. Precup, Revisiting heterophily for graph neural networks, arXiv preprint arXiv:2210.07606.

[16] S. Abu-El-Haija, B. Perozzi, A. Kapoor, N. Alipourfard, K. Lerman, H. Harutyunyan, G. Ver Steeg, A. Galstyan, Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing, in: international conference on machine learning, PMLR, 2019, pp. 21–29.

[17] T. Wang, D. Jin, R. Wang, D. He, Y. Huang, Powerful graph convolutional networks with adaptive propagation mechanism for homophily and heterophily, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 36, 2022, pp. 4210–4218.

[18] M. Li, A. Micheli, Y. G. Wang, S. Pan, P. Lió, G. S. Gnecco, M. Sanguineti, Guest editorial: deep neural networks for graphs: theory, models, algorithms, and applications, IEEE Transactions on Neural Networks and Learning Systems 35 (4) (2024) 4367–4372.

[19] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907.

[20] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Advances in neural information processing systems 30.

[21] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903.

[22] M. Chen, Z. Wei, Z. Huang, B. Ding, Y. Li, Simple and deep graph convolutional networks, in: International conference on machine learning, PMLR, 2020, pp. 1725–1735.

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

[24] J. Zhu, Y. Yan, M. Heimann, L. Zhao, L. Akoglu, D. Koutra, Heterophily and graph neural networks: Past, present and future, IEEE Data Engineering Bulletin.

[25] D. Jin, Z. Yu, C. Huo, R. Wang, X. Wang, D. He, J. Han, Universal graph convolutional networks, Advances in Neural Information Processing Systems 34 (2021) 10654–10664.

[26] W. Jin, T. Derr, Y. Wang, Y. Ma, Z. Liu, J. Tang, Node similarity preserving graph convolutional networks, in: Proceedings of the 14th ACM international conference on web search and data mining, 2021, pp. 148–156.

[27] S. Suresh, V. Budde, J. Neville, P. Li, J. Ma, Breaking the limit of graph neural networks by improving the assortativity of graphs with local mixing patterns, arXiv preprint arXiv:2106.06586.

[28] S. Li, D. Kim, Q. Wang, Restructuring graph for higher homophily via adaptive spectral clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 8622–8630.

[29] Y. Yan, M. Hashemi, K. Swersky, Y. Yang, D. Koutra, Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks, in: 2022 IEEE International Conference on Data Mining (ICDM), IEEE, 2022, pp. 1287–1292.

[30] J. Li, R. Zheng, H. Feng, M. Li, X. Zhuang, Permutation equivariant graph framelets for heterophilous graph learning, IEEE Transactions on Neural Networks and Learning Systems.

[31] Y. Dong, K. Ding, B. Jalaian, S. Ji, J. Li, Adagnn: Graph neural networks with adaptive frequency response filter, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 392–401.

[32] Z. Wu, S. Pan, G. Long, J. Jiang, C. Zhang, Beyond low-pass filtering: Graph convolutional networks with automatic filtering, IEEE Transactions on Knowledge and Data Engineering.

[33] Y. Wu, L. Hu, Y. Wang, Signed attention based graph neural network for graphs with heterophily, Neurocomputing 557 (2023) 126731.

[34] W. Duan, J. Xuan, M. Qiao, J. Lu, Learning from the dark: boosting graph convolutional neural networks with diverse negative samples, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 6550–6558.

[35] W. Duan, J. Lu, Y. G. Wang, J. Xuan, Layer-diverse negative sampling for graph neural networks, arXiv preprint arXiv:2403.11408.

[36] L. Du, X. Shi, Q. Fu, X. Ma, H. Liu, S. Han, D. Zhang, Gbk-gnn: Gated bi-kernel graph neural networks for modeling both homophily and heterophily, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 1550–1558.

[37] X. Li, R. Zhu, Y. Cheng, C. Shan, S. Luo, D. Li, W. Qian, Finding global homophily in graph neural networks when meeting heterophily, in: International Conference on Machine Learning, PMLR, 2022, pp. 13242–13256.

[38] K. Huang, Y. G. Wang, M. Li, et al., How universal polynomial bases enhance spectral graph neural networks: Heterophily, over-smoothing, and over-squashing, arXiv preprint arXiv:2405.12474.

[39] Y. Liu, Y. Zheng, D. Zhang, V. C. Lee, S. Pan, Beyond smoothing: Unsupervised graph representation learning with edge heterophily discriminating, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 37, 2023, pp. 4516–4524.

[40] Y. Choi, J. Choi, T. Ko, H. Byun, C.-K. Kim, Finding heterophilic neighbors via confidence-based subgraph matching for semi-supervised node classification, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 283–292.

[41] P. Li, Y. Wang, H. Wang, J. Leskovec, Distance encoding: Design provably more powerful neural networks for graph representation learning, Advances in Neural Information Processing Systems 33 (2020) 4465–4478.

[42] Y. Lu, J. Chen, C. Sun, J. Hu, Graph inference representation: Learning graph positional embeddings with anchor path encoding, arXiv preprint arXiv:2105.03821.

[43] J. You, R. Ying, J. Leskovec, Position-aware graph neural networks, in: International conference on machine learning, PMLR, 2019, pp. 7134–7143.

[44] J. Li, P. Zhou, C. Xiong, S. C. Hoi, Prototypical contrastive learning of unsupervised representations, arXiv preprint arXiv:2005.04966.

[45] Z. Yang, W. Cohen, R. Salakhudinov, Revisiting semi-supervised learning with graph embeddings, in: International conference on machine learning, PMLR, 2016, pp. 40–48.

[46] E. Chien, J. Peng, P. Li, O. Milenkovic, Adaptive universal generalized pagerank graph neural network, arXiv preprint arXiv:2006.07988.

[47] D. Lim, F. Hohne, X. Li, S. L. Huang, V. Gupta, O. Bhalerao, S. N. Lim, Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods, Advances in Neural Information Processing Systems 34 (2021) 20887–20902.

[48] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[49] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, ArXiv abs/1808.06670.
URL `https://api.semanticscholar.org/CorpusID:52055130`

[50] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (11).