

# Hierarchical Graph Pooling with Structure Learning

Zhen Zhang<sup>1</sup>, Jiajun Bu<sup>1</sup>, Martin Ester<sup>2</sup>, Jianfeng Zhang<sup>3</sup>,  
Chengwei Yao<sup>1</sup>, Zhi Yu<sup>1</sup>, Can Wang<sup>1</sup>

<sup>1</sup>Zhejiang University, <sup>2</sup>Simon Fraser University, <sup>3</sup>Alibaba Group  
zhen\_zhang@zju.edu.cn

## Abstract

Graph Neural Networks (GNNs), which generalize deep neural networks to graph-structured data, have drawn considerable attention and achieved state-of-the-art performance in numerous graph related tasks. However, existing GNN models mainly focus on designing graph convolution operations. The graph pooling (or downsampling) operations, that play an important role in learning hierarchical representations, are usually overlooked. In this paper, we propose a novel graph pooling operator, called Hierarchical Graph Pooling with Structure Learning (HGP-SL), which can be integrated into various graph neural network architectures. HGP-SL incorporates graph pooling and structure learning into a unified module to generate hierarchical representations of graphs. More specifically, the graph pooling operation adaptively selects a subset of nodes to form an induced subgraph for the subsequent layers. To preserve the integrity of graph’s topological information, we further introduce a structure learning mechanism to learn a refined graph structure for the pooled graph at each layer. By combining HGP-SL operator with graph neural networks, we perform graph level representation learning with focus on graph classification task. Experimental results on six widely used benchmarks demonstrate the effectiveness of our proposed model.

## Introduction

Deep neural networks with convolution and pooling layers have achieved great success in various challenging tasks, ranging from computer vision (He et al. 2016), natural language understanding (Bahdanau, Cho, and Bengio 2015) to video processing (Karpathy et al. 2014). The data in these tasks are typically represented in the Euclidean space (i.e., modeled as 2-D or 3-D tensors), thus usually containing locality and order information for the convolution operations (Defferrard, Bresson, and Vandergheynst 2016). However, in many real-world problems, a large amount of data, such as social networks, chemical molecules and biological networks, are lying on non-Euclidean domains that can be naturally represented as graphs. Due to the neural network’s powerful capabilities, it’s quite appealing to generalize the convolution and pooling operations to graph-structured data.

Recently, there have been a myriad of attempts to generalize the convolution operations to arbitrary graphs, referred to as graph neural networks (GNNs for short). In general, these algorithms can be classified into two big categories: spectral and spatial approaches. For the spectral methods, they typically define the graph convolution operations based on graph Fourier transform (Bruna et al. 2013; Defferrard, Bresson, and Vandergheynst 2016; Kipf and Welling 2017). For the spatial methods, the graph convolution operations are devised by aggregating the node representations directly from its neighborhood (Hamilton, Ying, and Leskovec 2017; Monti et al. 2017; Veličković et al. 2018; Morris et al. 2019). Majority of the aforementioned methods mainly involve transforming, propagating and aggregating node features across the graph, which can fit in the message passing scheme (Gilmer et al. 2017). GNNs have been applied to different types of graphs (Veličković et al. 2018; Derr, Ma, and Tang 2018), and obtained outstanding performance in numerous graph related tasks, including node classification (Kipf and Welling 2017), link prediction (Schlichtkrull et al. 2018; Zhang et al. 2018b) and recommendation (Ying et al. 2018a), *etc.*

Nevertheless, the pooling operations in graphs have not been extensively studied yet, though they act a pivotal part in learning hierarchical representations for the task of graph classification (Ying et al. 2018b). The goal of graph classification is to predict the label associated with the entire graph by utilizing its node features and graph structure information, i.e., a graph level representation is needed. GNNs are originally designed to learn meaningful node level representations, thus a commonly adopted approach to generate graph level representation is to globally summarize all the node representations in the graph. Although workable, the graph level representation generated via this way is inherently “flat”, since the entire graph structure information is neglected during this process. Furthermore, GNNs can only pass messages between nodes through edges, but cannot aggregate node information in a hierarchical way. Meanwhile, graphs often have different substructures and nodes are of different roles, therefore they should contribute differently to the graph level representation. For example, in the protein-protein interaction graphs, the certain substructures may rep-

resent some specific functionalities, which are of great significance to predict the whole graph characteristics. To capture both the graph’s local and global structure information, a hierarchical pooling process is demanded.

There exists some very recent work that focuses on the hierarchical pooling procedure in GNNs (Ying et al. 2018b; Gao and Ji 2019; Diehl 2019; Gao, Chen, and Ji 2019). These models usually coarsen the graphs through grouping or sampling nodes into subgraphs level by level, thus the entire graph information is gradually reduced to the hierarchical induced subgraphs. However, the graph pooling operations still have room for improvement. In node grouping approaches, the hierarchical pooling methods (Ying et al. 2018b; Diehl 2019) suffer from high computational complexity, which require additional neural networks to downsize the nodes. In node sampling approaches, the generated induced subgraph (Gao and Ji 2019; Lee, Lee, and Kang 2019) might fail to preserve the key substructures and eventually lose the completeness of graph topological information. For instance, two nodes that are not directly connected but sharing many common neighbors in the original graph might become unreachable from each other in the induced subgraph, even if intuitively they ought to be “close” in the subgraph. Therefore, the distorted graph structure will hinder the message passing in subsequent layers.

To address the aforementioned limitations, we propose a novel graph pooling operator HGP-SL to learn hierarchical graph level representations. Specifically, HGP-SL first adaptively selects a subset of nodes according to our **defined node information score**, which fully utilizes both the node features and graph topological information. In addition, the proposed graph pooling operation is a non-parametric step, therefore no additional parameters need to be optimized during this procedure. Then, we apply a structure learning mechanism with sparse attention (Martins and Astudillo 2016) to the pooled graph, aiming to learn a refined graph structure that preserves the key substructures in the original graph. We integrate the pooling operator into graph convolutional neural network to perform graph classification and the whole procedure can be optimized in an end-to-end manner. To summarize, the main contributions of this paper are as follows:

- We introduce a novel graph pooling operator HGP-SL that can be integrated into various graph neural network architectures. Similarly to the pooling operations in convolutional neural networks, our proposed graph pooling operation is non-parametric<sup>1</sup> and very easy to implement.
- To the best of our knowledge, we are the first to design a structure learning mechanism for the pooled graph, which has the advantage of learning a refined graph structure to preserve the graph’s key substructures.
- We conduct extensive experiments on six public datasets to demonstrate HGP-SL’s effectiveness as well as superiority compared to a range of state-of-the-art methods.

<sup>1</sup>Note that the pooling process itself is non-parametric, however the structure learning mechanism indeed has an attention parameter. Thus, the overall HGP-SL operator is not non-parametric.

## Related Work

### Graph Neural Networks

GNNs can be generally categorized into two branches: spectral and spatial approaches. The spectral methods typically define the parameterized filters according to graph spectral theory. (Bruna et al. 2013) first proposed to define convolution operations for graph in the Fourier transform domain. Due to its heavy computation cost, it has difficulty in scaling to large graphs. Later on, (Defferrard, Bresson, and Vandergheynst 2016) improved its efficiency by approximating the K-polynomial filters through Chebyshev expansion. GCN (Kipf and Welling 2017) further simplified the ChebNet by truncating the Chebyshev polynomial to the first-order approximation of the localized spectral filters.

The spatial approaches design convolution operations by directly aggregating the node’s neighborhood information. Among them, GraphSAGE (Hamilton, Ying, and Leskovec 2017) proposed an inductive algorithm that can generalize to unseen nodes by aggregating its neighborhood content information. GAT (Veličković et al. 2018) utilized attention mechanism to aggregate nodes’ neighborhood representations with different weights. JK-Net (Xu et al. 2018) leveraged flexible neighborhood ranges to enable better node representations. More details can be found in several comprehensive surveys on graph neural networks (Zhou et al. 2018; Zhang, Cui, and Zhu 2018; Wu et al. 2019). Nevertheless, the above mentioned two branches of GNNs are mainly designed for learning meaningful node representations, and unable to generate hierarchical graph representations due to the lack of pooling operations.

### Graph Pooling

Pooling operations in GNNs can scale down the size of inputs and enlarge the receptive fields, thus giving rise to better generalization and performance. DiffPool (Ying et al. 2018b) proposed to softly assign nodes to a set of clusters using neural networks, which forms a dense cluster assignment matrix and is computation expensive. gPool (Gao and Ji 2019) and SAGPool (Lee, Lee, and Kang 2019) devised a top-K node selection procedure to form an induced subgraph for the next input layer. Though efficient, it might lose the completeness of the graph structure information and result in isolated subgraphs, which will hamper the message passing process in subsequent layers. EdgePool (Diehl 2019) designed pooling operation by contracting the edges in the graph, but its flexibility is poor because it will always pool roughly half of the total nodes. iPool (Gao, Xiong, and Frossard 2019) presented a parameter-free pooling scheme which is invariant to graph isomorphism. EigenPool (Ma et al. 2019) introduced a pooling operator based on the graph Fourier transform, which controls the pooling ratio through spectral clustering and it’s also very time consuming.

In addition, there are also some approaches that perform global pooling. For instance, Set2Set (Vinyals, Bengio, and Kudlur 2015) implemented the global pooling operation by aggregating information through LSTMs (Hochreiter and Schmidhuber 1997). DGCNN (Zhang et al. 2018a) pooled the graph according to the last channel of the feature map

values which are sorted in the descending order. Graph topological based pooling operations are proposed in (Defferrard, Bresson, and Vandergheynst 2016) and (Rhee, Seo, and Kim 2017) as well, where Graclus method (Dhillon, Guan, and Kulis 2007) is employed as a pooling module.

## The Proposed Model

### Notations and Problem Formulation

Given a set of graph data  $G = \{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n\}$ , where the number of nodes and edges in each graph might be quite different. For an arbitrary graph  $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathbf{X}_i)$ , we have  $n_i$  and  $e_i$  denote the number of nodes and edges, respectively. Let  $\mathbf{A}_i \in \mathbb{R}^{n_i \times n_i}$  be the adjacent matrix describing its edge connection information and  $\mathbf{X}_i \in \mathbb{R}^{n_i \times f}$  represents the node feature matrix, where  $f$  is the dimension of node attributes. Label matrix  $\mathbf{Y} \in \mathbb{R}^{n \times c}$  indicates the associated labels for each graph, i.e., if  $\mathcal{G}_i$  belongs to class  $j$ , then  $\mathbf{Y}_{ij} = 1$ , otherwise  $\mathbf{Y}_{ij} = 0$ . Since the graph structure and node numbers change between layers due to the graph pooling operation, we further represent the  $i$ -th graph fed into the  $k$ -th layer as  $\mathcal{G}_i^k$  with  $n_i^k$  nodes. The adjacent matrix and hidden representation matrix are then denoted as  $\mathbf{A}_i^k \in \mathbb{R}^{n_i^k \times n_i^k}$  and  $\mathbf{H}_i^k \in \mathbb{R}^{n_i^k \times d}$ . With the above notations, we formally define our problem as follows:

**Input:** Given a set of graphs  $G_L$  with its label information  $\mathbf{Y}_L$ , the number of graph neural network layers  $K$ , pooling ratio  $r$ , and representation dimension  $d$  in each layer.

**Output:** Our goal is to predict the unknown graph labels of  $G/G_L$  with graph neural network in an end-to-end way.

### Graph Convolutional Neural Network

Graph convolutional neural network (or GCN) (Kipf and Welling 2017) has shown to be very efficient and achieved promising performance in various challenging tasks. Thus, we choose GCN as our model’s building block and briefly review its mechanism in this subsection. Please note that our proposed HGP-SL operator can also be integrated into other graph neural network architectures like GraphSAGE (Hamilton, Ying, and Leskovec 2017) and GAT (Veličković et al. 2018). We will discuss this in the experiment section. For the  $k$ -th layer in GCN, it takes graph  $\mathcal{G}$ ’s adjacent matrix  $\mathbf{A}$  and hidden representation matrix  $\mathbf{H}_k$  as input, then the next layer’s output will be generated as follows:

$$\mathbf{H}_{k+1} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}_k \mathbf{W}^k), \quad (1)$$

where  $\sigma(\cdot)$  is the non-linear activation function and  $\mathbf{H}_0 = \mathbf{X}$ ,  $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$  is the adjacent matrix with self-connections.  $\tilde{\mathbf{D}}$  is the diagonal degree matrix of  $\tilde{\mathbf{A}}$ , and  $\mathbf{W}^k \in \mathbb{R}^{d_k \times d_{k+1}}$  is a trainable weight matrix. For the ease of parameter tuning, we set output dimension  $d_{k+1} = d_k = d$  for all layers.

### The Overall Neural Network Architecture

Figure 1 provides an overview of our proposed Hierarchical Graph Pooling with Structure Learning (HGP-SL) that combines with graph neural network, where graph pooling operations are added between graph convolution operations. The

proposed HGP-SL operator is composed of two major components: 1) graph pooling, which preserves a subset of informative nodes and forms a smaller induced subgraph; and 2) structure learning, which learns a refined graph structure for the pooled subgraph. The advantage of our proposed structure learning lies in its capability to preserve the essential graph structure information, which will facilitate the message passing procedure. As in this illustrative example, the pooled subgraph might exist isolated nodes but intuitively ought to be connected, thus it would hinder the information propagation in subsequent layers especially when aggregating information from its neighborhood nodes. The whole architecture is the stacking of convolution and pooling operations, thus making it possible to learn graph representations in a hierarchical way. Then, a readout function is utilized to summarize node representations in each level, and the final graph level representation is the addition of different levels’ summarizations. At last, the graph level representation is fed into a Multi-Layer Perceptron (MLP) with softmax layer to perform graph classification task. In what follows, we give the details of graph pooling and structure learning layers.

### Graph Pooling Operation

In this subsection, we introduce our proposed graph pooling operation to enable down-sampling on graph data. Inspired by (Gao and Ji 2019; Lee, Lee, and Kang 2019; Gao, Xiong, and Frossard 2019), the pooling operation identifies a subset of informative nodes to form a new but smaller graph. Here, we design a non-parametric pooling operation, which can fully utilize both the node features and graph structure information.

The key of our proposed graph pooling operation is to define a criterion that guides the node selection procedure. To perform node sampling, we first introduce a criterion named **node information score** to evaluate the information that each node contains given its neighborhood. Generally, if a node’s representation can be reconstructed by its neighborhood representations, it means this node can probably be deleted in the pooled graph with almost no information loss. Here, we formally define the node information score as the Manhattan distance between the node representation itself and the one constructed from its neighbors:

$$\mathbf{p} = \gamma(\mathcal{G}_i) = \|(\mathbf{I}_i^k - (\mathbf{D}_i^k)^{-1} \mathbf{A}_i^k) \mathbf{H}_i^k\|_1, \quad (2)$$

where  $\mathbf{A}_i^k \in \mathbb{R}^{n_i^k \times n_i^k}$  and  $\mathbf{H}_i^k \in \mathbb{R}^{n_i^k \times d}$  are the adjacent and node representations matrices.  $\|\cdot\|_1$  performs  $\ell_1$  norm row-wisely.  $\mathbf{D}_i^k$  represents the diagonal degree matrix of  $\mathbf{A}_i^k$ , and  $\mathbf{I}_i^k$  is the identity matrix. Therefore, we have  $\mathbf{p} \in \mathbb{R}^{n_i}$  encode the information score of each node in the graph.

After having obtained the node information score, we can now select nodes that should be preserved by the pooling operator. To approximate the graph information, we choose to preserve the nodes that can not be well represented by their neighbors, i.e., the nodes with relative larger node information score will be preserved in the construction of the pooled graph, because they can provide more information. In details, we first re-order the nodes in graph according to their node information scores, then a subset of top-ranked

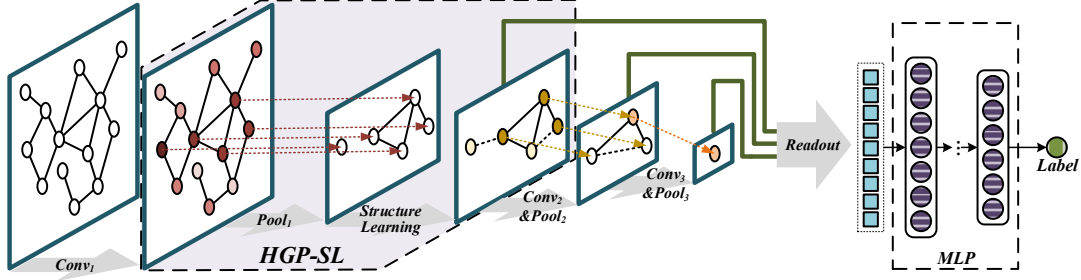


Figure 1: Architecture of proposed HGP-SL operator combined with graph neural network. The dashed box demonstrates the workflow of HGP-SL, which involves graph pooling and structure learning. The learned edges are represented as dashed lines in the graph. This procedure (convolution and pooling operations) is repeated several times. Then, a readout function is applied to aggregate node representations to make a fixed size representation, which goes through MLP layers for graph classification.

nodes are selected as follows:

$$\begin{aligned} \text{idx} &= \text{top-rank}(\mathbf{p}, \lceil r * n_i^k \rceil) \\ \tilde{\mathbf{H}}_i^{k+1} &= \mathbf{H}_i^k(\text{idx}, :) \\ \mathbf{A}_i^{k+1} &= \mathbf{A}_i^k(\text{idx}, \text{idx}), \end{aligned} \quad (3)$$

where  $r$  is the pooling ratio and  $\text{top-rank}(\cdot)$  denotes the function that returns the indices of the top  $n_i^{k+1} = \lceil r * n_i^k \rceil$  values.  $\mathbf{H}_i^k(\text{idx}, :)$  and  $\mathbf{A}_i^k(\text{idx}, \text{idx})$  perform the row or (and) column extraction to form the node representation matrix and adjacent matrix for the induced subgraph. Thus, we have  $\tilde{\mathbf{H}}_i^{k+1} \in \mathbb{R}^{n_i^{k+1} \times d}$  and  $\mathbf{A}_i^{k+1} \in \mathbb{R}^{n_i^{k+1} \times n_i^{k+1}}$  represent the node feature and graph structure information of next layer.

### Structure Learning Mechanism

In this subsection, we present how our proposed structure learning mechanism learns a refined graph structure in the pooled graph. As we have illustrated in Figure 1, the pooling operation might result in highly related nodes being disconnected in the induced subgraph, which loses the completeness of the graph structure information and further hinders the message passing procedure. Meanwhile, the graph structure obtained from domain knowledge (e.g., social network) or established by human (e.g., KNN graph) are usually non-optimal for the learning task in graph neural networks, due to the lost or noisy information. To overcome this problem, (Li et al. 2018) proposed to adaptively estimate graph Laplacian using an approximate distance metric learning algorithm, which might lead to local optimal solution. (Jiang et al. 2019) introduced to learn the constructed graph structure for node label estimation, however it generates dense connected graph and is not applicable in our hierarchical graph level representation learning scenario.

Here, we develop a novel structure learning layer, which learns sparse graph structure through sparse attention mechanism (Martins and Astudillo 2016). For graph  $\mathcal{G}_i$ 's pooled subgraph  $\mathcal{G}_i^k$  at its  $k$ -th layer, we take its structure information  $\mathbf{A}_i^k \in \mathbb{R}^{n_i^k \times n_i^k}$  and hidden representations  $\mathbf{H}_i^k \in \mathbb{R}^{n_i^k \times d}$  as input. Our target is to learn a refined graph structure that encodes the underlying pairwise relationship between each pair of nodes. Formally, we utilize a single layer neural network parameterized by a weight vector  $\vec{\mathbf{a}} \in \mathbb{R}^{1 \times 2d}$ . Then,

the similarity score between node  $v_p$  and  $v_q$  calculated by the attention mechanism can be expressed as:

$$\mathbf{E}_i^k(p, q) = \sigma(\vec{\mathbf{a}} [\mathbf{H}_i^k(p, :)] \|\mathbf{H}_i^k(q, :)]^\top) + \lambda \cdot \mathbf{A}_i^k(p, q), \quad (4)$$

where  $\sigma(\cdot)$  is the activation function like  $\text{ReLU}(\cdot)$  and  $\|$  represents the concatenation operation.  $\mathbf{H}_i^k(p, :) \in \mathbb{R}^{1 \times d}$  and  $\mathbf{H}_i^k(q, :) \in \mathbb{R}^{1 \times d}$  indicate the  $p$ -th and  $q$ -th row of matrix  $\mathbf{H}_i^k$ , which denote the representations of node  $v_p$  and  $v_q$ , respectively. Specifically,  $\mathbf{A}_i^k$  encodes the induced subgraph structure information, where  $\mathbf{A}_i^k(p, q) = 0$  if node  $v_p$  and  $v_q$  are not directly connected. We incorporate  $\mathbf{A}_i^k$  into our structure learning layer to bias the attention mechanism to give a relatively larger similarity score between directly connected nodes, and at the same time try to learn the underlying pairwise relationships between disconnected nodes.  $\lambda$  is a trade-off parameter between them.

To make the similarity score easily comparable across different nodes, we could normalize them across nodes using the softmax function:

$$\mathbf{S}_i^k(p, q) = \frac{\exp(\mathbf{E}_i^k(p, q))}{\sum_{m=1}^{n_i^k} \exp(\mathbf{E}_i^k(p, m))}. \quad (5)$$

However, the softmax transformation always has non-zero values and thus results in dense fully connected graph, which may introduce lots of noise into the learned structure. Hence, we propose to utilize sparsemax function (Martins and Astudillo 2016), which retains most the important properties of softmax function and has in addition the ability of producing sparse distributions. The sparsemax( $\cdot$ ) function aims to return the Euclidean projection of input onto the probability simplex and can be formulated as follows:

$$\begin{aligned} \mathbf{S}_i^k(p, q) &= \text{sparsemax}(\mathbf{E}_i^k(p, q)) \\ \text{sparsemax}(\mathbf{E}_i^k(p, q)) &= [\mathbf{E}_i^k(p, q) - \tau(\mathbf{E}_i^k(p, :))]_+, \end{aligned} \quad (6)$$

where  $[x]_+ = \max\{0, x\}$ , and  $\tau(\cdot)$  is the threshold function that returns a threshold according to the procedure shown in Algorithm 1. Thus,  $\text{sparsemax}(\cdot)$  preserves the values above the threshold and the other values will be truncated to zeros, which brings sparse graph structure. Similarly to softmax function,  $\text{sparsemax}(\cdot)$  also has the properties of



non-negative and sum-to-one, that's to say,  $\mathbf{S}_i^k(p, q) \geq 0$  and  $\sum_{q=1}^{n_i^k} \mathbf{S}_i^k(p, q) = 1$ . The proof procedure is available in the supplemental material.

---

**Algorithm 1** The calculation procedure of function  $\tau(\cdot)$

---

**Require:** input vector  $\mathbf{z} \in \mathbb{R}^n$ .

- 1: Sort  $\mathbf{z}$  into  $\mathbf{u}$ :  $u_1 \geq u_2 \geq \dots \geq u_n$ .
  - 2: Get  $\rho = \max\{1 \leq j \leq n : u_j + \frac{1}{j}(1 - \sum_{i=1}^j u_i) > 0\}$ .
  - 3: Define  $\tau(\mathbf{z}) = \frac{1}{\rho}(\sum_{i=1}^{\rho} u_i - 1)$ .
- 

## Improving Structure Learning Efficiency

For large scale graphs, it will be computation expensive to calculate the similarities between each pair of nodes during the learning of structure  $\mathbf{S}_i^k$ . If we further take graph's localization and smoothness properties into account, it is reasonable to constrain the calculation process within the node's  $h$ -hop neighborhood ( $h = 2$  or  $3$ ). Therefore, the computation cost of  $\mathbf{S}_i^k$  can be greatly reduced.

## GCN and Graph Pooling Revisiting

After having obtained the refined graph structure  $\mathbf{S}_i^k$ , we conduct graph convolution and pooling operations in the following layers based on  $\tilde{\mathbf{H}}_i^k$  and  $\mathbf{S}_i^k$  (instead of  $\mathbf{A}_i^k$ ). Thus, Equation (1) can be simplified as follows:

$$\mathbf{H}_i^k = \sigma(\mathbf{S}_i^k \tilde{\mathbf{H}}_i^k \mathbf{W}^k). \quad (7)$$

Since the learned  $\mathbf{S}_i^k$  satisfies  $\sum_{q=1}^{n_i^k} \mathbf{S}_i^k(p, q) = 1$ , therefore we have the diagonal matrix  $\mathbf{D}_i^k = \text{Diag}(d_1, d_2, \dots, d_{n_i^k})$  with  $d_p = \sum_{q=1}^{n_i^k} \mathbf{S}_i^k(p, q)$ , which degenerates to identity matrix  $\mathbf{I}_i^k$ . Similarly, the calculation of node information score in Equation (2) can also be simplified as below:

$$\mathbf{p} = \gamma(\mathcal{G}_i) = \|(\mathbf{I}_i^k - \mathbf{S}_i^k)\mathbf{H}_i^k\|_1, \quad (8)$$

which makes our model very easy to implement.

## The Readout Function and Output Layer

As we have demonstrated in Figure 1, the neural network architecture repeats the graph convolution and pooling operations for several times, thus we would observe multiple subgraphs with different size in each level:  $\mathbf{H}_i^1, \mathbf{H}_i^2, \dots, \mathbf{H}_i^K$ . To generate a fixed size graph level representation, we devise a readout function that aggregates all the node representations in the subgraph. Here, we simply use the concatenation of mean-pooling and max-pooling in each subgraph as follows:

$$\mathbf{r}_i^k = \mathcal{R}(\mathbf{H}_i^k) = \sigma\left(\frac{1}{n_i^k} \sum_{p=1}^{n_i^k} \mathbf{H}_i^k(p, :) \parallel \max_{q=1}^d \mathbf{H}_i^k(:, q)\right), \quad (9)$$

where  $\sigma(\cdot)$  is a nonlinear activation function and  $\mathbf{r}_i^k \in \mathbb{R}^{2d}$ . We then add<sup>2</sup> the readout outputs of different levels to form

---

<sup>2</sup>In our experiment, we use fixed size node representation across all layers, i.e.,  $d_k = \dots = d_1 = d = 128$ .

Datasets	# $ \mathcal{G} $	# $ \mathcal{V} $	Avg. $ \mathcal{V} $	Avg. $ \mathcal{E} $	# $ c $
ENZYMES	600	19,580	32.63	62.14	6
PROTEINS	1,113	43,471	39.06	72.82	2
D&D	1,178	334,925	284.32	715.66	2
NCI1	4,110	122,747	29.87	32.30	2
NCI109	4,127	122,494	29.68	32.13	2
Mutagenicity	4,337	131,488	30.32	30.77	2

Table 1: Statistics of the datasets.

our final graph level representation:

$$\mathbf{z}_i = \mathbf{r}_i^1 + \mathbf{r}_i^2 + \dots + \mathbf{r}_i^K, \quad (10)$$

which summarizes different levels' graph representations.

Finally, we feed the graph level representation into MLP layer with softmax classifier, and the loss function is defined as the cross-entropy of predictions over the labels:

$$\begin{aligned} \hat{\mathbf{Y}} &= \text{softmax}(\text{MLP}(\mathbf{Z})) \\ \mathcal{L} &= - \sum_{i \in L} \sum_{j=1}^c \mathbf{Y}_{ij} \log \hat{\mathbf{Y}}_{ij}, \end{aligned} \quad (11)$$

where  $\hat{\mathbf{Y}}_{ij}$  represents the predicted probability that graph  $\mathcal{G}_i$  belongs to class  $j$ , and  $\mathbf{Y}_{ij}$  is the ground truth.  $L$  denotes the training set of graphs that have labels.

## Experiments and Analysis

### Datasets

We adopt six commonly used public benchmarks<sup>3</sup> for empirical studies. Statistics of the six datasets are summarized in Table 1 with more descriptions as follows: **ENZYMES** (Borgwardt et al. 2005) is a dataset of protein tertiary structures, and each enzyme belongs to one of the 6 EC top-level classes. **PROTEINS** and **D&D** (Dobson and Doig 2003) are two protein graph datasets, where nodes represent the amino acids and two nodes are connected by an edge if they are less than 6 Angstroms apart. The label indicates whether or not a protein is a non-enzyme. **NCI1** and **NCI109** (Shervashidze et al. 2011) are two biological datasets screened for activity against non-small cell lung cancer and ovarian cancer cell lines, where each graph is a chemical compound with nodes and edges representing atoms and chemical bonds, respectively. **Mutagenicity** (Kazius, McGuire, and Bursi 2005) is a chemical compound dataset of drugs, which can be categorized into two classes: mutagen and non-mutagen.

### Baselines

**Graph Kernel Methods.** This group of methods perform graph classification by utilizing carefully designed kernels. We choose three classical algorithms: GRAPHLET (Shervashidze et al. 2009), Shortest-Path Kernel (SP) (Borgwardt and Kriegel 2005) and Weisfeiler-Lehman Kernel (WL) (Shervashidze et al. 2011) as baselines.

---

<sup>3</sup>Benchmarks are publicly available at <https://ls11-www.cs.tu-dortmund.de/staff/morris/graphkerneldatasets>

Categories	Baselines	ENZYMES	PROTEINS	D&D	NCI1	NCI109	Mutagenicity
Kernels	GRAPHLET	29.16 $\pm$ 5.63	72.23 $\pm$ 4.49	72.54 $\pm$ 3.83	62.48 $\pm$ 2.11	60.96 $\pm$ 2.37	56.65 $\pm$ 1.74
	SP	42.66 $\pm$ 5.38	75.71 $\pm$ 2.73	78.72 $\pm$ 3.89	67.44 $\pm$ 2.76	67.72 $\pm$ 2.28	71.63 $\pm$ 2.19
	WL	51.16 $\pm$ 6.19	76.16 $\pm$ 3.99	76.44 $\pm$ 2.35	76.65 $\pm$ 1.99	76.19 $\pm$ 2.45	80.32 $\pm$ 1.71
GNNs	GCN	43.66 $\pm$ 3.39	75.17 $\pm$ 3.63	73.26 $\pm$ 4.46	76.29 $\pm$ 1.79	75.91 $\pm$ 1.84	79.81 $\pm$ 1.58
	GraphSAGE	37.99 $\pm$ 3.71	74.01 $\pm$ 4.27	75.78 $\pm$ 3.91	74.73 $\pm$ 1.34	74.17 $\pm$ 2.89	78.75 $\pm$ 1.18
	GAT	39.83 $\pm$ 3.68	74.72 $\pm$ 4.01	77.30 $\pm$ 3.68	74.90 $\pm$ 1.72	75.81 $\pm$ 2.68	78.89 $\pm$ 2.05
Pooling	Set2Set	33.16 $\pm$ 3.21	79.33 $\pm$ 0.84	70.83 $\pm$ 0.84	69.62 $\pm$ 1.32	73.66 $\pm$ 1.69	80.84 $\pm$ 0.67
	DGCNN	32.16 $\pm$ 3.87	79.99 $\pm$ 0.44	70.06 $\pm$ 1.21	74.08 $\pm$ 2.19	78.23 $\pm$ 1.31	80.41 $\pm$ 1.02
	DiffPool	60.61 $\pm$ 3.94	79.90 $\pm$ 2.95	78.61 $\pm$ 1.32	77.73 $\pm$ 0.83	77.13 $\pm$ 1.49	80.78 $\pm$ 1.12
	EigenPool	63.97 $\pm$ 2.51	78.84 $\pm$ 1.06	78.63 $\pm$ 1.36	77.24 $\pm$ 0.96	75.99 $\pm$ 1.42	80.11 $\pm$ 0.73
	gPool	43.33 $\pm$ 2.88	80.71 $\pm$ 1.75	77.02 $\pm$ 1.32	76.25 $\pm$ 1.39	76.61 $\pm$ 1.39	80.30 $\pm$ 1.54
	SAGPool	43.99 $\pm$ 4.23	81.72 $\pm$ 2.19	78.70 $\pm$ 2.29	77.88 $\pm$ 1.59	75.74 $\pm$ 1.47	79.72 $\pm$ 0.79
	EdgePool	65.33 $\pm$ 4.36	82.38 $\pm$ 0.82	79.20 $\pm$ 2.61	76.56 $\pm$ 1.01	79.02 $\pm$ 1.89	81.41 $\pm$ 0.88
Proposed	HGP-SL <sub>NSL</sub>	60.18 $\pm$ 2.43	81.51 $\pm$ 1.69	77.24 $\pm$ 1.09	76.33 $\pm$ 1.43	76.32 $\pm$ 1.22	79.42 $\pm$ 0.58
	HGP-SL <sub>HOP</sub>	62.16 $\pm$ 2.11	83.03 $\pm$ 1.74	78.42 $\pm$ 1.37	77.72 $\pm$ 1.54	78.78 $\pm$ 1.09	79.88 $\pm$ 1.09
	HGP-SL <sub>DEN</sub>	63.51 $\pm$ 2.64	83.12 $\pm$ 0.84	78.11 $\pm$ 1.35	77.42 $\pm$ 1.23	78.76 $\pm$ 0.61	81.07 $\pm$ 1.02
	HGP-SL	<b>68.79 <math>\pm</math> 2.11</b>	<b>84.91 <math>\pm</math> 1.62</b>	<b>80.96 <math>\pm</math> 1.26</b>	<b>78.45 <math>\pm</math> 0.77</b>	<b>80.67 <math>\pm</math> 1.16</b>	<b>82.15 <math>\pm</math> 0.58</b>

Table 2: Graph classification in terms of accuracy with standard deviation (in percentage). We use **bold** to highlight wins.

**Graph Neural Networks.** Approaches in this group include representative graph neural networks: GCN (Kipf and Welling 2017), GraphSAGE (Hamilton, Ying, and Leskovec 2017) and GAT (Veličković et al. 2018), which are designed to learn meaningful node level representations. Therefore, we employ our proposed readout function to summarize the node representations for graph classification.

**Graph Pooling Models.** In this group, we further consider numerous models that combine GNNs with pooling operator for graph level representation learning. Set2Set (Vinyals, Bengio, and Kudlur 2015) and DGCNN (Zhang et al. 2018a) are two novel global graph pooling algorithms. Another five hierarchical graph pooling models including DiffPool (Ying et al. 2018b), gPool (Gao and Ji 2019), SAGPool (Lee, Lee, and Kang 2019), EdgePool (Diehl 2019) and EigenPool (Ma et al. 2019) are also compared as baselines.

**HGP-SL Variants.** To further analyze the effectiveness of our proposed HGP-SL operator, we consider four variants here: HGP-SL<sub>NSL</sub> (No Structure Learning) which discards the structure learning layer to verify the effectiveness of our proposed structure learning module, HGP-SL<sub>HOP</sub> which removes the structure learning layer and connects the nodes within its  $h$ -hops, HGP-SL<sub>DEN</sub> (DENse) which employs the structure learning layer to learn a dense graph structure with softmax function defined in Equation (5) and HGP-SL which utilizes sparsemax function define in Equation (6) to learn a sparse graph structure. Both HGP-SL<sub>DEN</sub> and HGP-SL use efficiency improved structure learning strategy.

**Experiment and Parameter Settings.** Following many previous work (Ying et al. 2018b; Ma et al. 2019), we randomly split each dataset into three parts: 80% as training set, 10% as validation set and the remaining 10% as test set. We repeat this randomly splitting process 10 times, and the average performance with standard derivation is reported. For baseline algorithms, we use the source code released by the

authors, and their hyper-parameters are tuned to be optimal based on the validation set. In order to ensure a fair comparison, the same neural network architectures are used for the existing pooling baselines and our proposed model. The dimension of node representations is set as 128 for all methods and datasets. We implement our proposed HGP-SL with PyTorch, and the Adam optimizer is utilized to optimize the model. The learning rate and weight decays are searched in  $\{0.1, 0.01, 0.001, 1e^{-4}, 1e^{-5}\}$ , pooling ratio  $r \in [0.1, 0.9]$  and layers  $K \in [1, 5]$ . The MLP consists of three fully connected layers with number of neurons in each layer setting as 256, 128, 64, followed by a softmax classifier. Early stopping criterion is employed in the training process, i.e., we stop training if the validation loss dose not decrease for 100 consecutive epochs. The source code is publicly available<sup>4</sup>.

## Performance on Graph Classification

The classification performance is reported in Table 2. To summarize, we have the following observations:

- First of all, a general observation we can draw from the results is that our proposed HGP-SL consistently outperforms other state-of-the-art baselines among all datasets. For instance, our method achieves about 3.08% improvement over the best baseline in PROTEINS dataset, which is 12.97% improvement over GCN with no hierarchical pooling mechanism. This verifies the necessity of adding graph pooling module.
- It is worth noting that the traditional graph kernel based methods demonstrate competitive performance. However, the carefully designed graph kernels typically involve massive human domain knowledge, which has difficulty in generalizing to graphs with arbitrary structures. Furthermore, the two-stage procedure of extracting graph features and performing graph classification might result in

<sup>4</sup>Code is available at <https://github.com/cszzhangzhen/HGP-SL>

sub-optimal performance.

- Being consistent with previous work’s findings (Ma et al. 2019), we also observe that the GNNs group can not achieve satisfied results. We argue that the major reason is because they ignore the graph structure information when globally summarizing the node representations, which further verifies the necessity of adding graph pooling module.
- In particular, the global pooling approaches Set2Set and DGCNN are surpassed by most of the hierarchical pooling methods with a few exceptions. This is because their learned graph representations are still “flat”, and the hierarchical structure information or functional units in the graph are ignored, which play an important role in predicting the entire graph labels.
- We note that the hierarchical pooling models can achieve relative better performance among most baselines, which further shows the effectiveness of the hierarchical pooling mechanism. Among them, gPool and SAGPool perform poorly in ENZYMES dataset. This may be due to the limited training samples per class resulting in the neural network overfitting. EdgePool gains superior performance in this group of competitors, which scales down the size of graphs by contracting each pair of nodes in the graph. Obviously, our proposed HGP-SL outperforms EdgePool with different gains for all settings.
- Finally, HGP-SL and HGP-SL<sub>DEN</sub> obtain better performance than HGP-SL<sub>NSL</sub> and HGP-SL<sub>HOP</sub>, which justifies the effectiveness of our proposed structure learning layer. Moreover, HGP-SL<sub>HOP</sub> performs worse than HGP-SL. This is because the disconnected nodes are still unreachable in its  $h$ -hops. HGP-SL further outperforms HGP-SL<sub>DEN</sub>, which indicates the learned dense graph structure might introduce additional noisy information and degenerate the performance. Furthermore, in the real-world scenario, graphs usually have sparse topologies, thus our proposed HGP-SL could learn more reasonable graph structures compared with HGP-SL<sub>DEN</sub>.

## Ablation Study and Visualization

### HGP-SL Convolutional Neural Network Architectures.

As mentioned in previous sections, our proposed HGP-SL can be integrated into various graph neural network architectures. We consider three most widely used graph convolutional architectures as our model’s building block to investigate the affect of different convolution operations: GCN (Kipf and Welling 2017), GraphSAGE (Hamilton, Ying, and Leskovec 2017) and GAT (Veličković et al. 2018). We evaluate them on three datasets, which cover both small and large datasets. Their results are shown in Table 3. Similar results can also be found in the remaining datasets, thus we omit them due to the limited space. As demonstrated in Table 3, the performance on graph classification varies depending on which dataset and the type of GNN in HGP-SL are chosen. In addition, we also combine the top-K selection procedure proposed in gPool and SAGPool with our proposed structure learning. We name them as gPool-SL and SAGPool-SL

Architectures	PROTEINS	NCI109	Mutagenicity
HGP-SL <sub>GCN</sub>	84.91±1.62	80.67±1.16	82.15±0.58
HGP-SL <sub>GAT</sub>	85.04±1.01	79.82±1.06	82.02±0.81
HGP-SL <sub>SAGE</sub>	84.99±0.82	80.11±0.96	81.96±0.97
gPool-SL	81.25±1.27	77.71±1.22	80.42±1.08
SAGPool-SL	82.67±1.42	78.01±1.50	80.00±1.22

Table 3: HGP-SL performance with different architectures.

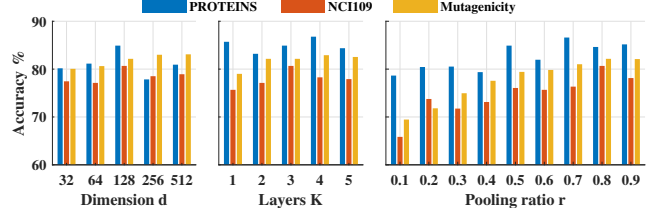


Figure 2: Hyper-parameter sensitivity analysis.

for short. From the results, we observe that gPool-SL and SAGPool-SL outperform gPool and SAGPool by incorporating the structure learning mechanism, which verifies the effectiveness of our proposed structure learning.

**Hyper-parameter Analysis.** We further study the sensitivities of several key hyper-parameters by varying them in different scales. Specifically, we investigate how the number of neural network layers  $K$ , graph representation dimension  $d$  and pooling ratio  $r$  will affect the graph classification performance. As we can see in Figure 2, HGP-SL almost achieves the best performance across different datasets when setting  $K = 3$ ,  $d = 128$  and  $r = 0.8$ , respectively. The pooling ratio  $r$  cannot be too small, otherwise most of the graph structure information will be lost during the pooling process.

**Visualization.** We utilize networkx<sup>5</sup> to visualize the pooling results of HGP-SL and its variants. In detail, we randomly sample a graph from PROTEINS dataset, which contains 154 nodes. We build a three layer graph neural network with pooling ratio setting as 0.5, which then generates three pooled graphs with nodes as 77, 39 and 20 respectively. We plot the 3rd pooled graph in Figure 3. It shows HGP-SL<sub>NSL</sub> and HGP-SL<sub>DEN</sub> fail to preserve meaningful graph topologies, while HGP-SL is able to preserve relatively reasonable topology of the original protein graph after pooling.

## Conclusion

In this paper, we investigate graph level representation learning for the task of graph classification. We propose a novel graph pooling operator HGP-SL, which empowers GNNs to learn hierarchical graph representations. It can also be conveniently integrated into various GNN architectures. Specifically, the graph pooling operation is a non-parametric step, which utilizes node features and graph structure information to perform down-sampling on graphs. Then, a structure learning layer is stacked on the pooling operation, which

<sup>5</sup><https://networkx.github.io/>

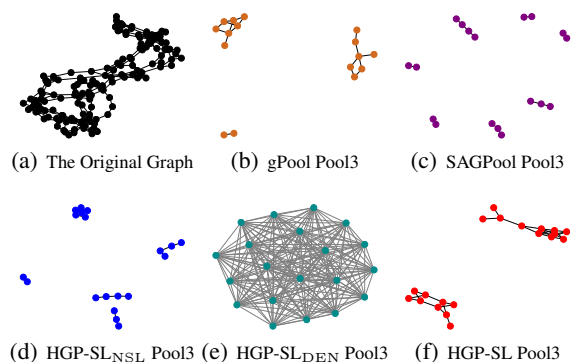


Figure 3: Visualization of different pooling methods.

aims to learn a refined graph structure that can best preserve the essential topological information. We combine the proposed HGP-SL operator with graph convolutional neural networks to conduct graph classification task. Comprehensive experiments on six widely used benchmarks demonstrate its superiority to a range of state-of-the-art methods.

## References

- [Bahdanau, Cho, and Bengio 2015] Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*.
- [Borgwardt and Kriegel 2005] Borgwardt, K. M., and Kriegel, H.-P. 2005. Shortest-path kernels on graphs. In *ICDM*. IEEE.
- [Borgwardt et al. 2005] Borgwardt, K. M.; Ong, C. S.; Schönauer, S.; Vishwanathan, S.; Smola, A. J.; and Kriegel, H.-P. 2005. Protein function prediction via graph kernels. *Bioinformatics* 47–56.
- [Bruna et al. 2013] Bruna, J.; Zaremba, W.; Szlam, A.; and LeCun, Y. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.
- [Defferrard, Bresson, and Vandergheynst 2016] Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, 3844–3852.
- [Derr, Ma, and Tang 2018] Derr, T.; Ma, Y.; and Tang, J. 2018. Signed graph convolutional networks. In *ICDM*, 929–934.
- [Dhillon, Guan, and Kulis 2007] Dhillon, I. S.; Guan, Y.; and Kulis, B. 2007. Weighted graph cuts without eigenvectors a multilevel approach. *TPAMI* 29(11):1944–1957.
- [Diehl 2019] Diehl, F. 2019. Edge contraction pooling for graph neural networks. *arXiv preprint arXiv:1905.10990*.
- [Dobson and Doig 2003] Dobson, P. D., and Doig, A. J. 2003. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology* 330(4):771–783.
- [Gao and Ji 2019] Gao, H., and Ji, S. 2019. Graph u-nets. In *International Conference on Machine Learning*, 2083–2092.
- [Gao, Chen, and Ji 2019] Gao, H.; Chen, Y.; and Ji, S. 2019. Learning graph pooling and hybrid convolutional operations for text representations. In *The World Wide Web Conference*, 2743–2749. ACM.
- [Gao, Xiong, and Frossard 2019] Gao, X.; Xiong, H.; and Frossard, P. 2019. ipool—information-based pooling in hierarchical graph neural networks. *arXiv preprint arXiv:1907.00832*.
- [Gilmer et al. 2017] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *ICML*, 1263–1272.
- [Hamilton, Ying, and Leskovec 2017] Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 1024–1034.
- [He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.
- [Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- [Jiang et al. 2019] Jiang, B.; Zhang, Z.; Lin, D.; Tang, J.; and Luo, B. 2019. Semi-supervised learning with graph learning-convolutional networks. In *CVPR*, 11313–11320.
- [Karpathy et al. 2014] Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*, 1725–1732.
- [Kazius, McGuire, and Bursi 2005] Kazius, J.; McGuire, R.; and Bursi, R. 2005. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of medicinal chemistry* 48(1):312–320.
- [Kipf and Welling 2017] Kipf, T. N., and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. *ICLR*.
- [Lee, Lee, and Kang 2019] Lee, J.; Lee, I.; and Kang, J. 2019. Self-attention graph pooling. In *ICML*, 3734–3743.
- [Li et al. 2018] Li, R.; Wang, S.; Zhu, F.; and Huang, J. 2018. Adaptive graph convolutional neural networks. In *AAAI*.
- [Ma et al. 2019] Ma, Y.; Wang, S.; Aggarwal, C. C.; and Tang, J. 2019. Graph convolutional networks with eigenpooling. In *SIGKDD*.
- [Martins and Astudillo 2016] Martins, A., and Astudillo, R. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *ICML*, 1614–1623.
- [Monti et al. 2017] Monti, F.; Boscaini, D.; Masci, J.; Rodola, E.; Svoboda, J.; and Bronstein, M. M. 2017. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR*, 5115–5124.
- [Morris et al. 2019] Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; and Grohe, M. 2019. Weisfeiler and leman go neural: Higher-order graph neural networks. In *AAAI*, 4602–4609.



- [Rhee, Seo, and Kim 2017] Rhee, S.; Seo, S.; and Kim, S. 2017. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. *arXiv preprint arXiv:1711.05859*.
- [Schlichtkrull et al. 2018] Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 593–607. Springer.
- [Shervashidze et al. 2009] Shervashidze, N.; Vishwanathan, S.; Petri, T.; Mehlhorn, K.; and Borgwardt, K. 2009. Efficient graphlet kernels for large graph comparison. In *AIS-TATS*, 488–495.
- [Shervashidze et al. 2011] Shervashidze, N.; Schweitzer, P.; Leeuwen, E. J. v.; Mehlhorn, K.; and Borgwardt, K. M. 2011. Weisfeiler-lehman graph kernels. *JMLR* 12(Sep):2539–2561.
- [Veličković et al. 2018] Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph attention networks. *ICLR*.
- [Vinyals, Bengio, and Kudlur 2015] Vinyals, O.; Bengio, S.; and Kudlur, M. 2015. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*.
- [Wu et al. 2019] Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Yu, P. S. 2019. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*.
- [Xu et al. 2018] Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; and Jegelka, S. 2018. Representation learning on graphs with jumping knowledge networks. *ICML*.
- [Ying et al. 2018a] Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018a. Graph convolutional neural networks for web-scale recommender systems. In *SIGKDD*, 974–983. ACM.
- [Ying et al. 2018b] Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018b. Hierarchical graph representation learning with differentiable pooling. In *NIPS*, 4800–4810.
- [Zhang et al. 2018a] Zhang, M.; Cui, Z.; Neumann, M.; and Chen, Y. 2018a. An end-to-end deep learning architecture for graph classification. In *AAAI*.
- [Zhang et al. 2018b] Zhang, Z.; Yang, H.; Bu, J.; Zhou, S.; Yu, P.; Zhang, J.; Ester, M.; and Wang, C. 2018b. Anrl: Attributed network representation learning via deep neural networks. In *IJCAI*, 3155–3161.
- [Zhang, Cui, and Zhu 2018] Zhang, Z.; Cui, P.; and Zhu, W. 2018. Deep learning on graphs: A survey. *arXiv preprint arXiv:1812.04202*.
- [Zhou et al. 2018] Zhou, J.; Cui, G.; Zhang, Z.; Yang, C.; Liu, Z.; and Sun, M. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*.

## Appendix

### Proof for Algorithm 1

To summarize,  $\text{sparsemax}(\cdot)$  considers the Euclidean projection of the input vector  $\mathbf{z}$  onto the probability simplex, which can be defined as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{p} \in \mathbb{R}^n} \quad & \frac{1}{2} \|\mathbf{p} - \mathbf{z}\|^2 \\ \text{s.t.} \quad & \mathbf{p}^\top \mathbf{1} = 1, \quad \mathbf{p} \geq 0. \end{aligned} \quad (12)$$

Then, the Lagrangian of the optimization problem in Equation (12) is:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\alpha}, \beta) = \frac{1}{2} \|\mathbf{p} - \mathbf{z}\|^2 - \boldsymbol{\alpha}^\top \mathbf{p} + \beta(\mathbf{1}^\top \mathbf{p} - 1). \quad (13)$$

The optimal  $(\mathbf{p}^*, \boldsymbol{\alpha}^*, \beta^*)$  must satisfy the following Karush-Kuhn-Tucker conditions:

$$\mathbf{p}^* - \mathbf{z} - \boldsymbol{\alpha}^* + \beta^* \mathbf{1} = \mathbf{0}, \quad (14)$$

$$\mathbf{1}^\top \mathbf{p}^* = 1, \quad \mathbf{p}^* \geq \mathbf{0}, \quad \boldsymbol{\alpha}^* \geq \mathbf{0}, \quad (15)$$

$$\alpha_i^* p_i^* = 0, \quad \forall i \in \{1, \dots, n\}. \quad (16)$$

If for  $\forall i \in \{1, \dots, n\}$  we have  $p_i^* > 0$ , then from Equation (16) we must satisfy  $\alpha_i^* = 0$ . Thus, from Equation (14) we can get  $p_i^* = z_i - \beta^*$ . Let  $S(\mathbf{z}) = \{j \in \{1, \dots, n\} | p_j^* > 0\}$ . From Equation (15) we obtain  $\sum_{j \in S(\mathbf{z})} (z_j - \beta^*) = 1$ , which yields the Line 3 in Algorithm 1, i.e.,  $\beta^* = \tau(\mathbf{z})$ . Again from Equation (16), we have that  $\alpha_i^* > 0$  implies  $p_i^* = 0$ , which from Equation (14) implies  $\alpha_i^* = \beta^* - z_i \geq 0$ , i.e.,  $z_i \leq \beta^*$  for  $i \notin S(\mathbf{z})$ . Thus, we have the procedure in Algorithm 1.