

Exploring Heterogeneous Information Networks via Pre-Training

1st Yang Fang
National University of Defense Technology
Changsha, China
fangyang12@nudt.edu.cn

2nd Xiang Zhao
National University of Defense Technology
Changsha, China
fangyang12@nudt.edu.cn

3rd Weidong Xiao
National University of Defense Technology
Changsha, China
wdxiao@nudt.edu.cn

Abstract—To explore heterogeneous information networks (HINs), network representation learning (NRL) is proposed, which represents a network in a low-dimension space. Recently, graph neural networks (GNNs) have drawn a lot of attention which are very expressive for mining a HIN, while they suffer from low efficiency issue. In this paper, we propose a pre-training and fine-tuning framework **PF-HIN** to capture the features of a HIN. Unlike traditional GNNs that have to train the whole model for each downstream task, **PF-HIN** only needs to fine-tune the model using the pre-trained parameters and minimal extra task-specific parameters, thus improving the model efficiency and effectiveness. Specifically, in pre-training phase, we first use a ranking-based BFS strategy to form the input node sequence. Then inspired by BERT, we adopt deep bi-directional transformer encoders to train the model, which is a variant of GNN aggregator that is more powerful than traditional deep neural networks like CNN and LSTM. The model is pre-trained based on two tasks, i.e., masked node modeling (MNM) and adjacent node prediction (ANP). Additionally, we leverage factorized embedding parameterization and cross-layer parameter sharing to reduce the parameters. In fine-tuning stage, we choose four benchmark downstream tasks, i.e., link prediction, similarity search, node classification and node clustering. We use node sequence pairs as input for link prediction and similarity search, and a single node sequence as input for node classification and clustering. The experimental results of the above tasks on four real-world datasets verify the advancement of **PF-HIN**, as it outperforms state-of-the-art alternatives consistently and significantly.

Index Terms—Heterogeneous information network; Bi-directional transformer; Pre-training

I. INTRODUCTION

In real world, complex systems are always related to multiple types of objects and relations. Such systems could be effectively abstracted via heterogeneous information networks (HINs), where different types of nodes (objects) are connected by unique edges (relations) [1]. Hence, compared with homogeneous networks that only possess a single type of nodes, HINs provide a richer tool to model the real-life issues.

In order to mine the abundant information behind the HIN, *network representation learning*, also known as network embedding learning, which embeds a network into

a low-dimensional space, has drawn lots of interests from researchers. Classical network embedding models like DeepWalk [2], LINE [3] and node2vec [4] are devised for homogeneous network using random walks to capture the structural information of networks. However, these methods are lacking in capability of expressing a HIN. Hence, models designed specifically for HINs have been proposed [5]–[7]. They are basically based on the *metapath*, which is a sequence of node types with edge types in between. To leverage the relationship between nodes and metapaths, different mechanisms are proposed, for example, heterogeneous SkipGram [5], proximity distance [6] and Hardmard function [7]. Nevertheless, these heterogeneous models’ performances confront the bottleneck due to the limited ability of metapath for capturing the features of a HIN.

Recently, graph neural networks (GNNs) have been investigated thoroughly, showing promising results on modeling the structural information of a network [8]–[10]. GNNs are usually empowered by complex encoders, basically deep neural networks like CNN, which could explore the neighborhood structure instead of a path, thus improving the performance on representing the HIN. However, to train such deep model on a HIN is often time-consuming and requires to train the whole model all over again for every specific task, leading to its inefficiency.

To address such issue, inspired by the recent advance in pre-training framework [11]–[13], we propose to pre-train our model on large datasets in the first place. And then for specific downstream task on specific dataset like DBLP, we use fine-tuning technique with minimal task-specific parameters, so as to improve the model efficiency and effectiveness. The above two-stage (**Pre-training** and **Fine-tuning**) framework for exploring the features of a **HIN** is named as **PF-HIN** in this paper.

In specific, in pre-training stage, inspired by BERT, we adopt deep bi-directional transformers to train the dataset. Thus we need to transform the node’s neighborhood into a sequence. We first measure the rankings of all nodes in HIN

based on their degree and betweenness centrality. Then we use ranking-based BFS strategy to generate the sequence, that is, always selecting the closest high-ranking nodes to form the sequence. Afterwards we prepare the input representation to be trained, which is the combination of token, segment, type, ranking and position embeddings. Note that in our paper, we use type embeddings to indicate the type information of a node.

During the pre-training stage, we adopt two strategies to reduce the parameters to further improve the model efficiency, i.e., factorized embedding parameterization and cross-layer parameter sharing. We design two tasks to pre-train PF-HIN. One is the masked node modeling (MNM) task, which is similar to masked language modeling mode. In this task, a certain percentage of nodes are masked and we need to predict those masked nodes. The other is the adjacent node prediction task which could capture the relationship between nodes. Given a node u_i having sequence X_i , our aim is to predict whether the node u_j with sequence X_j is the adjacent node. Notice that the operation which applies deep bi-directional transformers on the node sequence is actually a variant of GNN aggregating method, as those transformer layers could be regarded as deep neural networks. We would verify that such bi-directional transformer layers outperform traditional deep neural networks like CNN, LSTM or attention mechanism in ablation analysis.

During the fine-tuning stage, we choose four benchmark downstream tasks, i.e, link prediction, similarity search, node classification and node clustering. In link prediction and similarity search, we use node sequence pairs as input, identifying whether there is a link between two nodes and measuring the similarity between two nodes, respectively. In node classification and node clustering tasks, we use one single node sequence as input, employing softmax layer for classification and k-means algorithm for clustering, respectively. Our model PF-HIN advances state of the art on these downstream tasks consistently and significantly. We further verify our model's efficiency against other alternatives trained by randomly updated initial parameters, as our pre-trained parameters could be directly applied on all tasks and all datasets.

Our major contribution could be summarized into four components:

- We propose a novel pre-training and fine-tuning framework PF-HIN, which provides a new angle to mine the information behind a HIN. Such framework improves the model's effectiveness and efficiency on downstream tasks compared with other state-of-the-art alternatives.
- We utilize masked node modeling and adjacent node prediction tasks to pre-train the model, which could fully express the relationship between nodes.
- We adopt the deep bi-directional transformer encoders to capture the structural features of a HIN, and such model architecture is actually a variant of GNN. Such layers of transformer are proved to be more effective than traditional deep neural networks like CNN, LSTM and attention mechanism.

- PF-HIN outperforms state of the art consistently and significantly on four benchmark downstream tasks, i.e., link prediction, similarity search, node classification and node clustering.

The rest of the paper is organized as follows. In Section II we introduce the related work, and then justify the intuitions of our method with its theoretical analysis in Section III. Next, we conduct experimental studies on downstream tasks along with ablation analysis in Section IV. Finally, we conclude our findings in Section V.

II. RELATED WORK

A. Network Representation Learning

Network representation learning (NRL) methods could be traced back to those dimensionality reduction techniques [14]–[17] which typically utilize the feature vectors of the nodes to construct the affinity graph and then calculate the eigenvectors of it. Graph factorization model [18] represents the graph as an adjacency matrix, and generates a low-dimensional representation of a graph via the matrix factorization. However, these models all suffer from the high computational cost and source data sparsity, and are unable to capture the global network structure [3].

Recently, many researches harness the power of random walks or paths in a network to help preserve the local and global structure of a network. DeepWalk [2] leverages random walks and applies SkipGram word2vec model to learn network embedding. node2vec [4] is actually an extension of DeepWalk, as it adopts a biased random walk strategy, which can better explore network structure. LINE [3] harnesses the first-order and second-order proximities simultaneously to encode local and neighborhood structure information. Some methods also utilize the text information behind a network to help learn the embeddings. For example, text-associated DeepWalk (TADW) [19] incorporates text information with the matrix factorization based DeepWalk. CANE [20] utilizes mutual-attention mechanism to learn context-aware network embedding. Max-margin DeepWalk (MMDW) [21] leverages nodes labeling information to learn discriminative network embedding. Group-enhanced network embedding (GENE) [22] integrates existing group information into NRL. Context-enhanced network embedding (CENE) [23] regards text content as a special kind of node, thus harnessing both textual and structural information to learn the representation.

While aforementioned approaches are designed for homogeneous network, there is also dedicated research specifically exploiting the features of heterogeneous network. PTE [24] defines the conditional probability of nodes of one type generated by nodes of another type, then makes the conditional distribution close to its empirical distribution. metapath2vec [5] proposes a heterogeneous SkipGram with its context window restricted to one specific type. HINE [6] proposes metapath based proximity, and preserves such proximity by minimizing the distance between nodes' joint probability defined by sigmoid and empirical probabilities. HIN2Vec [7] devises

Hadamard multiplication of nodes and metapaths to capture features of a HIN. UNRA [25] simultaneously preserves inter-relationship among homogeneous nodes and node-content correlation, and relationships between different types of nodes are also learned and assembled in a unified framework.

B. Graph Neural Network

In recent, lots of graph neural network (GNN) models have been proposed, showing promising results on representing network data. Here we summarize the latest development of GNN.

Inspired by the success of convolutional neural network (CNN) in the computer vision domain, many efforts have been dedicated to generalizing such convolutional operation from traditional data like images to graph data. [26] is the first prominent work to propose a spectral graph theory based graph convolution operation, in which given a node, its neighborhood is considered as a receptive field to be aggregated recursively. GCN [8] adopts localized first-order approximation of spectral graph convolutions to improve the scalability of the model. There is a line of research to further improve the spectral based GNN models [27]–[30], however, it processes the whole graph simultaneously, thus leading to its inefficiency. To alleviate the problem, the spatial-based GNN models emerges rapidly [31]–[34]. For example, GraphSAGE [34] leverages a sampling strategy to sample the neighboring nodes iteratively, instead of the whole graph. LGCN [33] utilizes a sub-graph training method to reduce the excessive memory and computational resource requirements.

In addition to those convolution based models, graph attention networks are proposed, which also intend to fuse the neighboring nodes or walks in graphs to learn a new node representation [9], [35], [36]. The major difference is that it introduces the attention mechanism to assign higher weights on more important nodes or walks. For example, GAT [9] harnesses the masked self-attentional layers to apply different weights to different nodes in a neighborhood, being more efficient on graph-structured data.

The above GNN models are all originally devised for homogeneous networks as they aggregate the neighboring nodes or walks regardless of their types. To make it more adaptive to HIN, HetGNN [10] first samples a fixed size of neighboring nodes of a given node and then groups those nodes based on their specific types. Afterwards, it uses a neural network architecture with two modules to aggregate the feature information of the above neighboring nodes. Specifically, one module is to encode features of each type of nodes, the other is to aggregate features of different types.

C. Pre-training Approach

In this paper, we borrow the idea of pre-training to help explore the HIN, so as to serve for the downstream tasks. Pre-trained approaches have been proved to be useful to many natural language tasks and here we introduce the recent development of pre-training.

ELMo extracts context-sensitive features via bi-directional LSTM, and uses task-specific architectures that include the pre-trained representations as additional features [12]. OpenAI GPT [37] leverages the left-to-right architecture in which every token can only attend to previous tokens in the self-attention layers of the transformer [38]. BERT [39] adopts a masked language model for pre-training which could fuse the left and right context, so that a multi-layer bi-directional transformer encoder could be applied. It also includes “next sentence prediction” task to jointly pre-train text-pair representations. However, BERT neglects dependency between the masked positions and suffers from a pretrain-finetune discrepancy. To overcome this challenge, XLNet [40] proposes a generalized autoregressive pretraining method which maximizes the expected likelihood over all permutations of the factorization order, so as to learn the bidirectional contexts. It overcomes BERT’s limitation via its autoregressive formulation. RoBERTa [41] dynamically changes the masking pattern applied on the training data. It also removes the next sentence prediction task and train with larger batches to improve the efficiency. Recently, ALBERT [42] is proposed to further improve the efficiency and effectiveness of such pretrain-finetune framework. Unlike BERT, it adopts factorized embedding parameterization to reduce the number of parameters. It also shares all the cross-layer parameters including both feed-forward network and attention parameters. Moreover, it introduces the sentence-order prediction (SOP) loss instead of next sentence prediction loss, which helps the model focus on the inter-sentence coherence.

III. PROPOSED MODEL

We first provide preliminaries, and then describe the node sequence generation procedure, the input representation, followed by the pre-training and fine-tuning stages of our model PF-HIN.

A. Preliminaries

First, we introduce the definitions of HIN. A HIN is a graph $G = (V, E, T)$, where V denotes the set of nodes and E denotes the set of edges between nodes. Each node and edge is associated with a type mapping function, $\phi : V \rightarrow T_V$ and $\varphi : E \rightarrow T_E$, respectively. T_V and T_E denote the sets of node and edge types. A HIN is a network where $|T_V| > 1$ and/or $|T_E| > 1$.

B. Node Sequence Generation

We first transform the structure of node’s neighbor to a sequence, so that deep bi-directional transformers could be applied. The maximum length of a node sequence is set to k . We use ranking-based breadth-first search (BFS) strategy to generate the sequence. Specifically, we use node degree and betweenness centrality to measure the ranking of a node in a HIN. Betweenness centrality is to calculate the fractions of shortest paths that pass through node u . Such ranking function could be represented as $r : V \rightarrow \{1, \dots, |V|\}$. The higher degree and betweenness centrality, the lower the ranking. Based

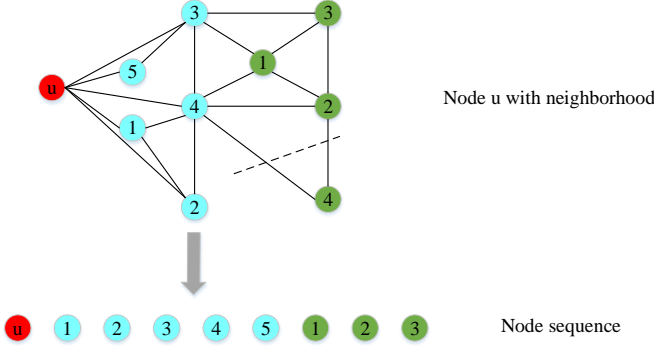


Fig. 1: The example of node sequence generation. Given a node u with $k = 9$, the blue nodes are the 1-hop nodes and the green nodes are the 2-hop nodes. The sequence first includes the 1-hop nodes arranged by node ranking and then includes the 2-hop nodes in the same way. Note that the maximum length of the sequence is 9, so we exclude the green node with ranking label 4.

on the ranking algorithm above, for each node, we mark it with a ranking label. We denote the labeling procedure as $l : V \rightarrow O$, which transfers the node set V to an ordered set O . For node u and node v , $l(u) < l(v)$ if and only if $r(u) < r(v)$. Afterwards, for node u , we first choose its 1-hop neighbor nodes and put them in an ordered sequence arranged by their ranking label. Then we choose 2-hop nodes to add them in the sequence in the same way. We iterate such operation until the sequence length is extended to k . Figure 1 illustrates the node sequence generation procedure harnessing the ranking-based BFS.

C. Input Representation

Similar to BERT, to enable PF-HIN to handle the downstream tasks, we allow our input representation to be able to unambiguously represent both a pair of node sequences and a single node sequence in one input sequence. The first token of the input sequence is always a special classification token ([CLS]) and the final hidden state of this token is calculated by aggregating the sequence representation. The node token is the name of a node, and we use the whole name of the node as one token. For two sequences packed together, we use a special token [SEP] to differentiate them, and we also add segment embedding to distinguish every node according to which sequence it belongs to.

In order to identify different node types in HIN, we add a learned embedding to input representation, indicating which type a node belongs to. Moreover, we transfer node ranking label as embeddings to indicate the ranking information. Then the overall input representation is constructed by summing token, segment, type, ranking and position embeddings. Figure 2 illustrates the above construction of the input representation.

D. Model Architecture

To increase the training speed of our model, we adopt two parameter reduction techniques to lower memory limitations

inspired by ALBERT. It shares the similar architecture of BERT using layers of bi-directional transformers. However, unlike BERT where the node embedding size E is tied with the hidden layer size H , i.e., $E \equiv H$, we make a more efficient usage of the total model parameters by untying E and H considering the modeling needs, which dictates that $H \gg E$. We adopt factorized embedding parameterization which decomposes the parameters into two smaller matrices. Rather than mapping the one-hot vectors directly to hidden space with size H , we first map them to a low-dimensional embedding space with size E , and then map it to the hidden space. The detailed information could be found in [42]. Afterwards we adopt cross-layer parameter sharing to further boost the efficiency. Traditional sharing mechanism either only shares the feed forward network (FFN) parameters across layers or only shares the attention parameters. Our default decision is to share all the parameters across layers.

We denote the number of transformer layers as L , and the number of self-attention heads as A . So the configuration of our model is that L is set to 12, H is set to 768, A is set to 12, E is set to 128, and the number of total parameters is equal to 12M.

E. Pre-training PF-HIN

In pre-training stage, we adopt two tasks to pre-train PF-HIN, i.e., masked node modeling and adjacent node prediction. Figure 3 presents the framework of the pre-training procedure.

1) *Masked Node Modeling*: Our masked node modeling task (MNM) resembles the masked LM model, and we randomly mask a certain percentage of the input nodes and then predict those masked nodes. Given input sequence x_1, x_2, \dots, x_n , we randomly choose 15% nodes to be replaced. And for a chosen node x_i , we replace its token with the actual [MASK] token with 80% probability, another random node token with 10% probability and the unchanged x_i with 10% probability. Afterwards the masked sequence is fed into the bi-directional transformer encoders. Such transformer encoder performs like an aggregator in GNN. The final hidden state h_i^L corresponding to the [MASK] token is fed to a feedforward layer. Then the output is used to predict the target node via a softmax classification layer:

$$z_i = \text{Feedforward}(h_i^L), \quad (1)$$

$$\mathbf{p}_i = \text{softmax}(\mathbf{W}^{\text{MNM}} z_i), \quad (2)$$

where z_i is the output of the feedforward layer, $\mathbf{W}^{\text{MNM}} \in V \times d$ is the classification weight shared with the input node embedding matrix, V is the number of nodes, d is the dimension of the hidden state size, \mathbf{p}_i is the predicted distribution of x_i over all nodes.

For training, we use cross-entropy between the one-hot label \mathbf{y}_i and the prediction \mathbf{p}_i :

$$\mathcal{L} = - \sum_t y_t \log p_t, \quad (3)$$

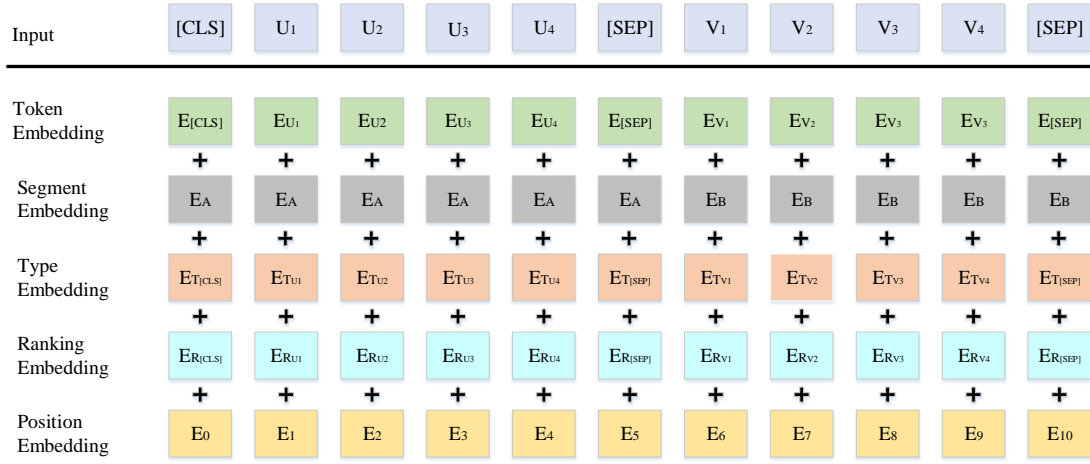


Fig. 2: The construction of input representation.

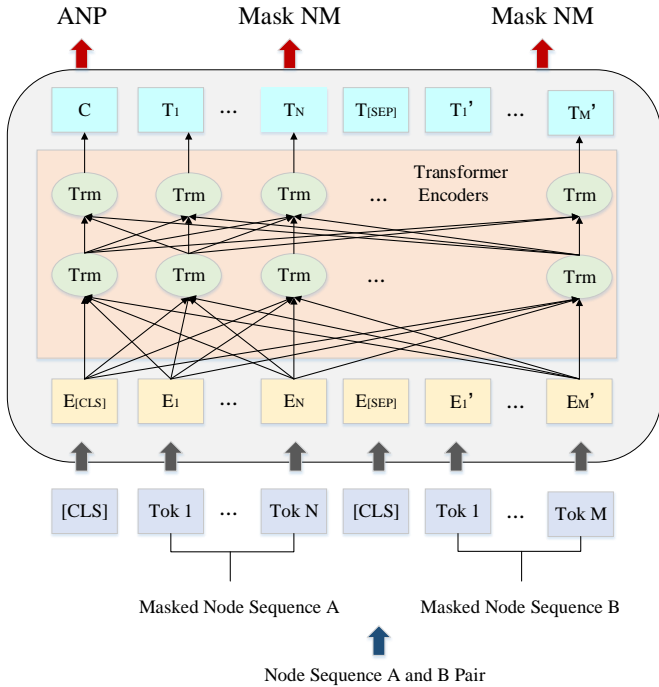


Fig. 3: The framework of pre-training. [CLS] is a special symbol for classification added in front of every input sequence, and [SEP] is a special separator token to separate the node sequence pair.

where y_t and p_t are the t -th components of \mathbf{y}_i and \mathbf{p}_i , respectively. Here we adopt a smoothing strategy by setting $y_t = \epsilon$ for the target node and $y_t = \frac{1-\epsilon}{V-1}$ for each of the other nodes. By doing so, we could lessen the restriction that one-hot label corresponds to only one answer.

2) *Adjacent Node Prediction*: Many downstream tasks are based on capturing the relationship between nodes, like link prediction and similarity search. We design a pre-training model to conduct a binarized adjacent node prediction task, so as to understand the node relationships. In specific, for node

u_i with sequence X_i and node u_j with sequence X_j , 50% of the time we choose u_j being the actual adjacent node of u_i (labeled as IsAdjacent), and 50% of the time we randomly choose u_j from the corpus (labeled as NotAdjacent). As shown in Figure 3, C is used for adjacent node prediction (ANP). Given the classification layer weights W^{ANP} , the scoring function s_τ of whether the node pair is adjacent is shown as follows:

$$s_\tau = \text{sigmoid}(CW^{ANP^T}), \quad (4)$$

in which $s_\tau \in \mathbb{R}^2$ is a binary vector with $s_{\tau 0}, s_{\tau 1} \in [0, 1]$ and $s_{\tau 0} + s_{\tau 1} = 1$.

Considering the positive adjacent node pair \mathbb{S}^+ and a negative adjacent node pair \mathbb{S}^- , we calculate a cross-entropy loss as follows:

$$\mathcal{L} = - \sum_{\tau \in \mathbb{S}^+ \cup \mathbb{S}^-} (y_\tau \log(s_{\tau 0}) + (1 - y_\tau) \log(s_{\tau 1})), \quad (5)$$

in which y_τ is the label (positive or negative) of that node pair.

F. Fine-tuning PF-HIN

The self-attention mechanism in the transformer allows PF-HIN to model many downstream tasks. Fine-tuning could be realized by simply swapping out the proper inputs and outputs, regardless of single node sequence or sequence pairs. For each downstream task, the task-specific inputs and outputs are simply plugged into PF-HIN and all the parameters are fine-tuned end-to-end. Here we introduce four tasks, i.e., link prediction, similarity search, node classification and node clustering.

Specifically, in link prediction, we aim to predict whether there is a link between two nodes, and the inputs are the node sequence pairs. For output, we feed the [CLS] representation into the sigmoid layer, so as to predict the existence of a link between two nodes. The only new parameters are the classification layer weights $W \in \mathbb{R}^2 \times H$, where H is the

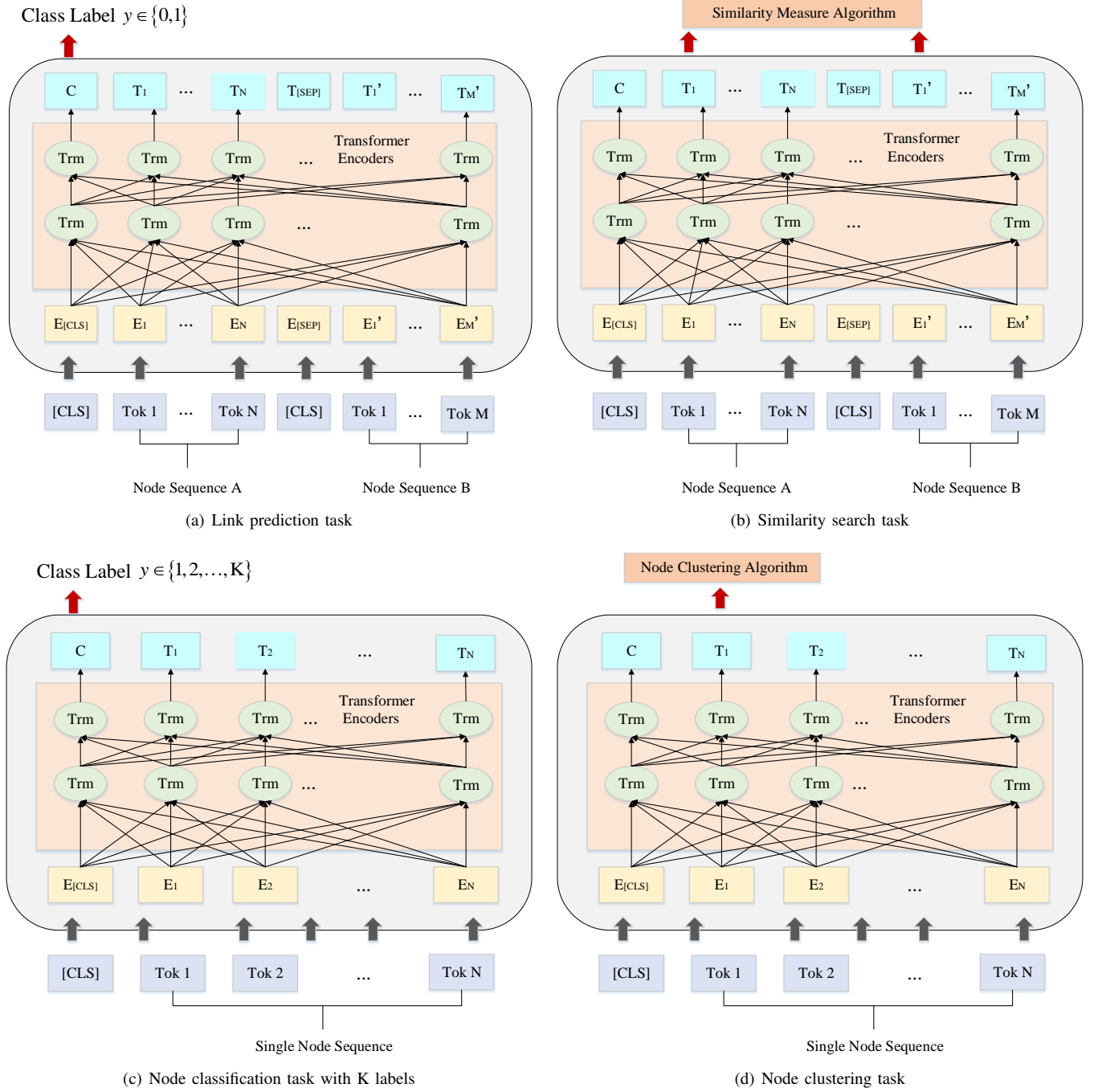


Fig. 4: Illustrations of fine-tuning PF-HIN on different tasks.

size of hidden state. Figure 4(a) illustrates the fine-tuning architecture of this task.

In similarity search, in order to measure the similarity between two nodes, we use the node sequence pairs as input. We leverage the token-level output representations to compute the similarity score of two nodes. See Figure 4(b) for the detailed information of the fine-tuning procedure.

In node classification, we only use a single node sequence as input and generate the classification label based on the [CLS]

representation. It is fed into a softmax layer, calculating the classification loss with the new parameters, i.e., classification layer weights $W \in \mathbb{R}^K \times H$, where K is the number of classification labels and H is the size of hidden state. Figure 4(c) shows the specific model structure.

In node clustering, we also use one node sequence as input and then put the token-level output embeddings to a clustering model, so as to cluster the data. Figure 4(d) presents the details of this fine-tuning framework.

The experimental information of the above tasks will be introduced in the corresponding subsections in Section IV.

IV. EXPERIMENTS

In this section, we first introduce the experimental setup and baseline models. Afterwards we present the PF-HIN fine-tuning results on four downstream tasks, i.e., link prediction, similarity search, node classification and node clustering, along with the computation cost on each task. We also conduct the ablation analysis and parameter sensitivity research.

A. Experimental setup

1) *Datasets*: We conduct experiments on four different datasets, DBLP,¹ YELP,² YAGO³ and Freebase.⁴ DBLP is a bibliographic dataset having four types of nodes, i.e., author, paper, venue and topic. The authors are split into four areas: machine learning, data mining, database and information retrieval. YELP is a social media dataset, consisting of reviews on restaurants. It also has four types of nodes, i.e., review, customer, restaurant and food-related keywords. The restaurants are separated as Chinese food, fast food and sushi bar. YAGO is a knowledge base and we extracted a subset of it containing movie information, having five types of nodes, i.e., movie, actor, director, composer and producer. The movies are split into five types, i.e., action, adventure, sci-fi, crime and horror. Freebase also contains lots of real-life knowledge and facts, and we extracted a subset of video games, containing four types of nodes, i.e., game, publisher, developer, designer. The video games are divided into three types, i.e., adventure, strategy and action.

To conduct the pre-training procedure, we use the combination of the above four datasets. Overall it contains about 10M nodes and 35M edges. And for the specific downstream tasks, we use much smaller training datasets which are in line with those having been reported in the existing literature. The training dataset statistics are presented in Table I.

TABLE I: Dataset statistics.

Dataset	#nodes	#edges	# node types	# labels
DBLP	301,273	1,382,587	4	4
YELP	201,374	872,432	4	3
YAGO	52,384	143,173	5	4
Freebase	42,374	122,364	4	3

The datasets extracted from DBLP and YELP are larger than YAGO and Freebase, and we aim to testify that PF-HIN is scalable to both small and large datasets.

2) *Compared Algorithms*: We first choose DeepWalk, LINE and node2vec as baselines, which were originally applied on homogeneous information networks. DeepWalk and node2vec all leverage random walks, while node2vec uses a biased walk strategy to better capture the network

structure. LINE explores the local and neighborhood structural information via first-order and second-order proximities.

We also include three state-of-the-art algorithms devised for HIN, i.e., metapath2vec, HINE, HIN2Vec. They are all based on meta-path, differing in their mechanisms of harnessing the meta-path features. Specifically, metapath2vec adopts a heterogeneous skipgram, HINE proposes a metapath-based notion of proximity and HIN2Vec utilizes the Hadamard multiplication of nodes and metapaths.

Our transformer operation on nodes could be regarded as a special aggregator in GNN model. For a fair comparison, we include several other GNN models, i.e., GCN, GAT and GraphSAGE, which were originally devised for homogeneous information network. GCN and GraphSAGE are based on convolutional operations, while GCN requires the full graph Laplacian and GraphSAGE only needs a node's local neighborhood. GAT employs the attention mechanism to capture the correlation between central node and neighboring nodes. We also select a GNN model designed for HIN, i.e., HetGNN, which samples the heterogeneous neighbors, grouping based on their node types, and then aggregate feature information of those sampled neighboring nodes.

3) *Parameters*: For pre-training, we set the generated sequence length k as 20. The dimension of node embedding is set to 128 and the size of hidden state is set to 768. On transformer layers, we use 0.1 as the dropout probability. The Adam learning rate is initiated as 0.001 with a linear decay. We use 256 sequences to form a batch and the training epoch is set to 20. The training loss is the sum of the mean masked node modeling likelihood and the mean adjacent node prediction likelihood.

In fine-tuning, most parameters remain the same as those in pre-training, except the learning rate, batch size and number of epochs. We leverage the grid search to set the best configuration. The learning rate is chosen from $\{0.01, 0.02, 0.025, 0.05\}$. The training epoch is chosen from $\{2, 3, 4, 5\}$. The batch size is chosen from $\{16, 32, 64\}$. The optimal parameters values are task-specific.

We report on statistical significance with a paired two-tailed t-test and we mark a significant improvement of PF-HIN over HetGNN for $p < 0.05$ with \blacktriangle .

B. Downstream Tasks

1) *Link Prediction*: This task is to predict which links would occur in the future. Unlike previous work [4] that randomly samples certain percentage of links as training dataset and uses the remaining as evaluation dataset, we adopt a sequential split of training and testing data. In specific, we first train a binary logistic classifier on the graph of training data, and then we use the testing dataset with the same number of random negative (non-existent) links to evaluate the trained classifier. Additionally, we only consider the new links in the training dataset and remove those duplicate links from evaluation. We adopt AUC and F1 scores as evaluation metrics.

We present the link prediction results in Table II, with the best results highlighted in bold. From it we could observe

¹<http://dblp.uni-trier.de>

²https://www.yelp.com/dataset_challenge

³<https://old.datahub.io/dataset/yago>

⁴<https://developers.google.com/freebase/>

TABLE II: Experimental results on the link prediction task.

Model	DBLP		YELP		YAGO		Freebase	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
DeepWalk	0.583	0.351	0.602	0.467	0.735	0.525	0.793	0.632
LINE	0.579	0.357	0.605	0.463	0.739	0.531	0.803	0.625
node2vec	0.584	0.355	0.609	0.471	0.742	0.534	0.801	0.635
metapath2vec	0.604	0.367	0.618	0.473	0.744	0.541	0.806	0.642
HINE	0.607	0.369	0.621	0.482	0.763	0.548	0.816	0.647
HIN2Vec	0.611	0.376	0.625	0.493	0.768	0.578	0.821	0.657
GCN	0.623	0.392	0.638	0.516	0.779	0.583	0.833	0.674
GraphSage	0.627	0.395	0.641	0.525	0.783	0.592	0.834	0.678
GAT	0.631	0.392	0.644	0.537	0.781	0.596	0.838	0.672
HetGNN	0.642	0.402	0.663	0.544	0.793	0.601	0.846	0.683
PF-HIN	0.649[▲]	0.407[▲]	0.671[▲]	0.551[▲]	0.804[▲]	0.612[▲]	0.852[▲]	0.686[▲]

that the outputs become better with a decreasing scale of the datasets. Traditional homogeneous models like DeepWalk, LINE and node2vec perform worse than traditional heterogeneous meta-path based models metapath2vec, HINE and HIN2Vec, which indicates that meta-path captures the network structure better than random walks. However, homogeneous GNN models like GCN, GraphSAGE and GAT have even better outputs comparing traditional heterogeneous methods. We attribute this to the fact that deep neural network explores the whole network in a better way, generating better representations for link prediction. HetGNN outperforms those homogeneous GNN models, since it takes the node types into consideration. Our model PF-HIN outperforms all the baselines consistently and significantly, which verifies that our fine-tuning framework based on bi-directional transformers is effective on modeling the relationships between nodes, so as to predict the links in between.

2) *Similarity Search*: In this task, we aim to find those similar nodes of a given node. In order to evaluate the similarity between two nodes, we directly calculate the cosine similarity based on the node representations. It is hard to rank all pairs of nodes explicitly, so we give an estimation based on the grouping label $g(\cdot)$, in which similar nodes are gathered in one group. Given a specific node u , if we rank other nodes based on the similarity score, intuitively, nodes from the same group (similar ones) should be at the top of the ranking list while those dissimilar ones should be ranked at the bottom. More specifically, we define the AUC value as follows.

$$AUC = \frac{1}{|V|} \sum_{u \in V} \frac{\sum_{v, v' \in V \wedge g(u)=g(v) \wedge g(u) \neq g(v')} \text{sim}(u, v) > \text{sim}(u, v')}{\sum_{v, v' \in V \wedge g(u)=g(v) \wedge g(u) \neq g(v')}}. \quad (6)$$

We train the models on the whole dataset while the AUC metric is computed only in the subset of nodes having grouping labels. The subset is relatively small since AUC value requires pairwise similarities among the subset.

Table III illustrates the experimental results of similarity search. The best results are highlighted in bold. According to this table, we could observe that traditional heterogeneous models and homogeneous GNN models achieve comparable outputs, which means that both meta-path based mechanism and deep neural networks can generate expressive node embeddings for similarity search. HetGNN is still the best baseline which proves the power of the combination of GNN

TABLE III: Experimental results on the similarity search task.

Model	DBLP	YELP	YAGO	Freebase
	AUC	AUC	AUC	AUC
DeepWalk	0.511	0.553	0.656	0.721
LINE	0.506	0.558	0.661	0.727
node2vec	0.513	0.559	0.653	0.731
metapath2vec	0.545	0.578	0.673	0.754
HINE	0.551	0.583	0.679	0.752
HIN2Vec	0.556	0.587	0.684	0.759
GCN	0.553	0.581	0.682	0.762
GraphSage	0.557	0.586	0.689	0.764
GAT	0.555	0.584	0.691	0.768
HetGNN	0.563	0.592	0.694	0.772
PF-HIN	0.569[▲]	0.601[▲]	0.707[▲]	0.783[▲]

and type features. PF-HIN performs the best in all cases, illustrating the effectiveness of our pre-training and fine-tuning framework on learning the node representations for similarity search.

3) *Node Classification*: Here we report on the experimental results for the multi-label node classification task. The classification labels of each dataset are introduced in Section IV-A1. We adopt micro-f1 (MIC-F1) and macro-f1 (MAC-F1) as evaluation metrics.

TABLE IV: Experimental results on the multi-label node classification task.

Model	DBLP		YELP		YAGO		Freebase	
	MIC-F1	MAC-F1	MIC-F1	MAC-F1	MIC-F1	MAC-F1	MIC-F1	MAC-F1
DeepWalk	0.193	0.191	0.163	0.145	0.328	0.265	0.541	0.480
LINE	0.184	0.179	0.274	0.276	0.366	0.320	0.514	0.447
node2vec	0.201	0.198	0.194	0.151	0.332	0.280	0.539	0.487
metapath2vec	0.209	0.207	0.264	0.269	0.370	0.332	0.514	0.434
HINE	0.234	0.230	0.276	0.284	0.401	0.363	0.519	0.434
HIN2Vec	0.246	0.241	0.291	0.306	0.428	0.394	0.561	0.503
GCN	0.257	0.256	0.302	0.311	0.459	0.447	0.569	0.511
GraphSage	0.267	0.269	0.305	0.318	0.464	0.456	0.571	0.528
GAT	0.271	0.273	0.303	0.315	0.469	0.462	0.578	0.533
HetGNN	0.285	0.282	0.309	0.321	0.478	0.471	0.583	0.539
PF-HIN	0.293[▲]	0.291[▲]	0.315[▲]	0.329[▲]	0.488[▲]	0.480[▲]	0.591[▲]	0.548[▲]

Table IV presents the results of node classification, where the best outputs are highlighted in bold. From which we could observe that GNN based models are the best baselines, showing the advancement of deep neural network on exploring the features of the network data for classification. Our PF-HIN still obtains the best experimental results thanks to our fine-tuning framework which aggregates the whole sequence information for node classification.

4) *Node Clustering*: In this section, we report on the outcomes of the node clustering task. We feed the generated node embeddings of each model into a clustering model. Here we choose k-means algorithm to cluster the data. We leverage the normalized mutual information (NMI) and adjusted rand index (ARI) as evaluation metrics.

Table V shows the performance on the node clustering task, with the best outputs highlighted in bold. According to this table, we find that despite the homogeneous GNN models' strong ability of capturing the structural information of a network, they still perform slightly worse than those traditional heterogeneous models. And we attribute this to

TABLE V: Experimental results on the node clustering task.

Model	DBLP		YELP		YAGO		Freebase	
	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
DeepWalk	0.672	0.686	0.713	0.744	0.856	0.886	0.884	0.911
LINE	0.678	0.693	0.705	0.739	0.861	0.894	0.880	0.905
node2vec	0.673	0.689	0.719	0.748	0.867	0.899	0.887	0.916
metapath2vec	0.711	0.738	0.748	0.785	0.896	0.917	0.918	0.944
HINE	0.718	0.741	0.753	0.795	0.899	0.921	0.910	0.947
HIN2Vec	0.721	0.744	0.739	0.779	0.902	0.923	0.914	0.936
GCN	0.701	0.719	0.744	0.775	0.881	0.907	0.903	0.926
GraphSage	0.705	0.722	0.746	0.778	0.885	0.911	0.906	0.929
GAT	0.709	0.728	0.748	0.782	0.893	0.915	0.909	0.931
HetGNN	0.729	0.748	0.759	0.788	0.904	0.926	0.919	0.955
PF-HIN	0.734[▲]	0.759[▲]	0.771	0.795[▲]	0.911[▲]	0.932[▲]	0.927[▲]	0.962[▲]

taking node type information into consideration, which could make a real difference when clustering the nodes. No doubt HetGNN is the best baseline combining both type information and deep neural networks. Our PF-HIN still outperforms all the baselines, proving that PF-HIN is able to generate effective node embeddings for node clustering.

C. Computation Cost

To evaluate the efficiency of our fine-tuning framework comparing other models, here we conduct the computation cost analysis. In specific, we first extract a subset of DBLP used for downstream task, i.e., 15% of the original DBLP dataset. Then in this subset, we analyze the running time of each model in each task, using the early stopping mechanism. For simplicity consideration, here we use LP, SS, MC and NC to represent link prediction, similarity search, multi-label node classification and node clustering tasks, respectively. The models are all conducted on GPU GTX-1080. The computation cost results are illustrated in Table VI.

TABLE VI: Computation cost on different tasks.

Model	LR	SS	MC	NC
	Time (s)	Time (s)	Time (s)	Time (s)
DeepWalk	573	623	398	291
LINE	637	712	403	335
node2vec	784	837	522	399
metapath2vec	900	1028	733	569
HINE	993	1138	729	588
HIN2Vec	1080	1367	836	682
GCN	1863	2235	1436	1124
GraphSage	1479	1735	1173	791
GAT	1367	1589	938	836
HetGNN	1673	1987	1366	974
PF-HIN	822	921	646	478

According to Table VI, among all models, PF-HIN's running time is only a bit longer than those three traditional homogeneous models, DeepWalk, LINE and node2vec which are based on random walks. However, they perform the worst among all models as illustrated in Section IV-B. GNN based models like GCN, GraphSAGE, GAT and HetGNN cost much more time than other models, since the complexity of traditional deep neural networks is much higher than other algorithms. Although PF-HIN employs the bi-directional

transformer encoders which is a variant of GNN, it is much more efficient than the above GNN based models. This is because that pre-trained parameters and embeddings could help the cost function converge much faster.

D. Ablation Analysis

In this section, we conduct ablation experiments to further analyze the importance of each component of PF-HIN. In specific, we analyze the effect of pre-training tasks, bi-directional transformer encoders, the components of the input representation and the ranking-based BFS sampling strategy.

1) *Effect of Pre-training Tasks*: To evaluate the effect of pre-training tasks, we introduce two variants of PF-HIN, i.e., PF-HIN-MNM and PF-HIN-ANP. PF-HIN-MNM is the model excluding pre-training the masked node modeling task, PF-HIN-ANP is the model excluding pre-training the adjacent node prediction task. We conduct the experiments only on DBLP.

TABLE VII: Ablation analysis over pre-training tasks.

Model	LR		SS	MC		NC	
	AUC	F1	AUC	MIC-F1	MAC-F1	NMI	ARI
PF-HIN	0.649	0.407	0.569	0.293	0.291	0.734	0.759
PF-HIN-ANP	0.467	0.321	0.365	0.278	0.273	0.693	0.704
PF-HIN-MNM	0.342	0.213	0.256	0.183	0.176	0.348	0.369

Table VII shows the experimental results of the ablation analysis over pre-training tasks. We have the following observations. The MNM task plays a more important role for pre-training the model, as the performance drops dramatically after removing MNM. PF-HIN-ANP is slightly worse than PF-HIN on node classification and clustering tasks, while the gap is much larger on link prediction and similarity search tasks. This is because that ANP has a larger influence on fine-tuning framework with sequence pairs, which models the relationships between nodes.

2) *Effect of Bi-Directional Transformer Encoder*: Note that our bi-directional transformer encoder is actually a variant of GNN applied on HIN, aggregating the neighborhood information. Here we intend to replace our transformer encoders with CNN, bi-directional LSTM and attention mechanism. In specific, the model using CNN encoder is denoted as PF-HIN +CNN, the model using bi-directional LSTM is denoted as PF-HIN +LSTM and the model using attention mechanism is denoted as PF-HIN +attention. Here we conduct the experiments on DBLP.

TABLE VIII: Ablation analysis over transformer encoders.

Model	LR		SS	MC		NC	
	AUC	F1	AUC	MIC-F1	MAC-F1	NMI	ARI
PF-HIN	0.649	0.407	0.569	0.293	0.291	0.734	0.759
PF-HIN +CNN	0.625	0.396	0.559	0.285	0.278	0.706	0.721
PF-HIN +LSTM	0.633	0.399	0.555	0.274	0.267	0.712	0.722
PF-HIN +attention	0.626	0.394	0.558	0.282	0.269	0.712	0.725

Table VIII presents the experimental results of different variants. We could observe that CNN, LSTM and attention mechanism based models achieve comparable results on four

tasks. However, PF-HIN consistently outperforms all the traditional deep neural network models, which further proves that our bi-directional transformer encoders’ advancement on mining the information behind a HIN.

3) *Effect of Input Representation*: As mentioned in Section III-C, our input representation is composed of token, segment, type, ranking and position embeddings. Comparing the input representations in BERT, we add type and rank embeddings. Here we conduct the ablation experiments to analyze the effect of type and rank embeddings. The model excluding type embeddings is denoted as PF-HIN-type; the model excluding rank embeddings is denoted as PF-HIN-rank; and the model excluding both type and rank embeddings is denoted as PF-HIN-rank-type.

TABLE IX: Ablation analysis over input representation.

Model	LR		SS	MC		NC	
	AUC	F1	AUC	MIC-F1	MAC-F1	NMI	ARI
PF-HIN	0.649	0.407	0.569	0.293	0.291	0.734	0.759
PF-HIN-type	0.625	0.387	0.542	0.278	0.269	0.643	0.674
PF-HIN-rank	0.635	0.397	0.549	0.284	0.281	0.726	0.734
PF-HIN-type-rank	0.604	0.363	0.521	0.246	0.241	0.623	0.656

We present the ablation experimental results of input representation in Table IX. Type and rank embeddings could help improve the model performance as PF-HIN obtains the best results on all tasks. Offering the model with a sense of type information has a large influence on the performance, since removing it deteriorates the results a lot, especially on node clustering task. This is because that node clustering task is more sensitive to the type information. Removing rank embeddings also hurts the performance, since the ranking information could help assign higher weights on important nodes in a HIN. No doubt that the model PF-HIN-type-rank achieves the worst performance.

4) *Effect of the Ranking-Based BFS Sampling Strategy*: In this paper, we adopt the ranking-based BFS sampling strategy to sample the nodes to form the input sequence. Here we introduce two variant models, one is to use only BFS sampling strategy, denoted as PF-HIN +BFS, the other is to randomly choose the neighboring nodes to form the node sequence, denoted as PF-HIN +random.

TABLE X: Ablation analysis over Ranking-Based BFS Sampling Strategy.

Model	LR		SS	MC		NC	
	AUC	F1	AUC	MIC-F1	MAC-F1	NMI	ARI
PF-HIN	0.649	0.407	0.569	0.293	0.291	0.734	0.759
PF-HIN +BFS	0.624	0.394	0.541	0.278	0.278	0.712	0.733
PF-HIN +random	0.612	0.378	0.534	0.264	0.261	0.641	0.662

Table X shows the ablation experimental results. PF-HIN +BFS outperforms PF-HIN +random, which is because that aggregating the node’s closest neighborhood’s information is more expressive than randomly chosen neighboring nodes. PF-HIN still outperforms PF-HIN +BFS, which indicates that choosing the nodes with higher importance is better for describing the features of a HIN.

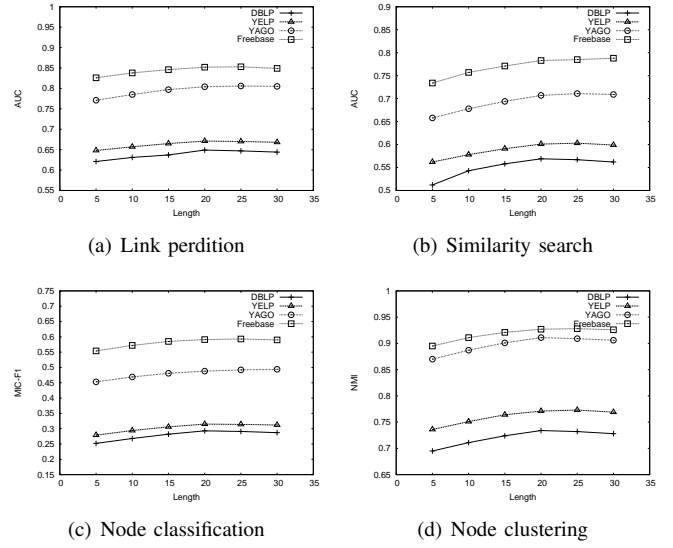


Fig. 5: Sensitivity analysis of maximum length of input sequence

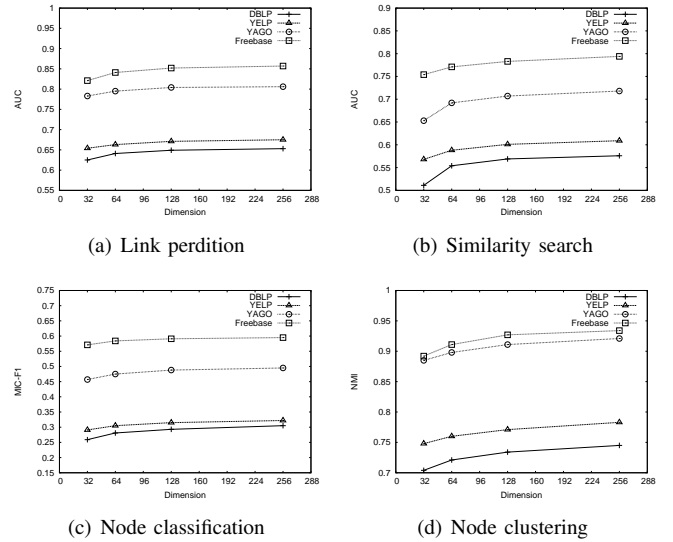


Fig. 6: Sensitivity analysis of dimension of node embeddings.

E. Parameter Sensitivity

Here we conduct the sensitivity analysis of hyperparameters. We choose two parameters to be analyzed, one is the maximum length of input sequence k , the other is the dimension of the node embedding E . For each task, we only choose one metric for evaluation. Specifically, we choose AUC value for link prediction, AUC value for similarity search, MIC-F1 value for node classification and NMI value for node clustering. Figure 5 and Figure 6 illustrate the experimental results of the parameter analysis.

As to the maximum length of input sequence, from Figure 5, we could observe that the performance improves rapidly when the length gets longer until it reaches 20. We attribute this to the fact that a short node sequence is not able to fully express

the neighborhood information. When the length reaches 20 or longer, the performance becomes steady and in some tasks and some datasets, longer sequence length even hurts the performance. For example, in Figure 5(d), on YAGO, the performance peaks at length 20 and becomes worse with the length getting longer. This is because that, given a node, its neighboring information could be well represented by its close neighborhood, however, including more far-away nodes may cause noise. According to this analysis, we choose the length of input sequence as 20 to balance the effectiveness and efficiency.

As to the dimension of node embeddings, we could observe from Figure 6 that, the performance becomes better with the dimension growing larger among all tasks and all datasets. This is because that higher dimension is able to capture more features. Basically, our model is not very sensitive to dimension, especially when the dimension is larger than 128. The performance gap is not very large between dimension 128 and 256. Thus we choose 128 as our experimental setup for efficiency consideration.

V. CONCLUSIONS

In this paper, we propose a novel model, namely, PF-HIN to mine the sufficient information behind a HIN. It is a pre-training and fine-tuning framework. In pre-training stage, we first adopt ranking-based BFS strategy to generate the input sequence. Then we leverage the bi-directional transformer layers to pre-train the model. We adopt factorized embedding parameterization and cross-layer parameter sharing strategies to reduce the parameters. The pre-training tasks we utilize are masked node modeling (MNM) and adjacent node prediction (ANP). Afterwards we fine-tune PF-HIN on four different tasks, i.e., link prediction, similarity search, node classification and node clustering. PF-HIN significantly and consistently outperforms baseline models on the above tasks on four real-life datasets.

In future work, it is of interest to see how to model a dynamic HIN that is constantly evolving, using a pre-training and fine-tuning framework.

ACKNOWLEDGMENT

This work was supported by NSFC under grants Nos. 61872446, 61902417, 71971212 and 71690233, and NSF of Hunan province under grant No. 2019JJ20024.

REFERENCES

- [1] Y. Sun and J. Han, *Mining Heterogeneous Information Networks: Principles and Methodologies*, ser. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2012. [Online]. Available: <https://doi.org/10.2200/S00433ED1V01Y201207DMK005>
- [2] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: online learning of social representations," in *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 2014, pp. 701–710. [Online]. Available: <https://doi.org/10.1145/2623330.2623732>
- [3] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, 2015, pp. 1067–1077. [Online]. Available: <https://doi.org/10.1145/2736277.2741093>
- [4] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 855–864. [Online]. Available: <https://doi.org/10.1145/2939672.2939754>
- [5] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, 2017, pp. 135–144. [Online]. Available: <https://doi.org/10.1145/3097983.3098036>
- [6] Z. Huang and N. Mamoulis, "Heterogeneous information network embedding for meta path based proximity," *CoRR*, vol. abs/1701.05291, 2017.
- [7] T. Fu, W. Lee, and Z. Lei, "Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, 2017, pp. 1797–1806. [Online]. Available: <https://doi.org/10.1145/3132847.3132953>
- [8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. [Online]. Available: <https://openreview.net/forum?id=SUJ4ayYgl>
- [9] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. [Online]. Available: <https://openreview.net/forum?id=JXmpikCZ>
- [10] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, 2019, pp. 793–803. [Online]. Available: <https://doi.org/10.1145/3292500.3330961>
- [11] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015, pp. 3079–3087. [Online]. Available: <http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning>
- [12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, 2018, pp. 2227–2237. [Online]. Available: <https://doi.org/10.18653/v1/n18-1202>
- [13] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 2018, pp. 328–339. [Online]. Available: <https://www.aclweb.org/anthology/P18-1031/>
- [14] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, 2001, pp. 585–591. [Online]. Available: <http://papers.nips.cc/paper/1961-laplacian-eigenmaps-and-spectral-techniques-for-embedding-and-clustering>
- [15] M. A. A. Cox and T. F. Cox, *Multidimensional Scaling*. Springer Berlin Heidelberg, 2008.
- [16] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [17] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1225–1234. [Online]. Available: <http://doi.acm.org/10.1145/2939672.2939753>
- [18] A. Ahmed, N. Shervashidze, S. M. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization," in *22nd International World Wide Web Conference, WWW '13, Rio de*

- Janeiro, Brazil, May 13-17, 2013, 2013, pp. 37–48. [Online]. Available: <https://doi.org/10.1145/2488388.2488393>
- [19] C. Yang, Z. Liu, D. Zhao, M. Sun, and E. Y. Chang, “Network representation learning with rich text information,” in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 2015, pp. 2111–2117.
 - [20] C. Tu, H. Liu, Z. Liu, and M. Sun, “CANE: context-aware network embedding for relation modeling,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 2017, pp. 1722–1731.
 - [21] C. Tu, W. Zhang, Z. Liu, and M. Sun, “Max-margin deepwalk: Discriminative learning of network representation,” in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 2016, pp. 3889–3895. [Online]. Available: <http://www.ijcai.org/Abstract/16/547>
 - [22] J. Chen, Q. Zhang, and X. Huang, “Incorporate group information to enhance network embedding,” in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, 2016, pp. 1901–1904. [Online]. Available: <https://doi.org/10.1145/2983323.2983869>
 - [23] X. Sun, J. Guo, X. Ding, and T. Liu, “A general framework for content-enhanced network representation learning,” *CoRR*, vol. abs/1610.02906, 2016.
 - [24] J. Tang, M. Qu, and Q. Mei, “PTE: predictive text embedding through large-scale heterogeneous text networks,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, 2015, pp. 1165–1174. [Online]. Available: <http://doi.acm.org/10.1145/2783258.2783307>
 - [25] R. Hu, C. P. Yu, S. Fung, S. Pan, H. Wang, and G. Long, “Universal network representation for heterogeneous information networks,” in *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, 2017, pp. 388–395. [Online]. Available: <https://doi.org/10.1109/IJCNN.2017.7965880>
 - [26] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, “Spectral networks and locally connected networks on graphs,” in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. [Online]. Available: <http://arxiv.org/abs/1312.6203>
 - [27] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 3837–3845. [Online]. Available: <http://papers.nips.cc/paper/6081-convolutional-neural-networks-on-graphs-with-fast-localized-spectral-filtering>
 - [28] M. Henaff, J. Bruna, and Y. LeCun, “Deep convolutional networks on graph-structured data,” *CoRR*, vol. abs/1506.05163, 2015. [Online]. Available: <http://arxiv.org/abs/1506.05163>
 - [29] R. Li, S. Wang, F. Zhu, and J. Huang, “Adaptive graph convolutional neural networks,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 3546–3553. [Online]. Available: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16642>
 - [30] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, “Cayleynets: Graph convolutional neural networks with complex rational spectral filters,” *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 97–109, 2019. [Online]. Available: <https://doi.org/10.1109/TSP.2018.2879624>
 - [31] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein, “Geometric deep learning on graphs and manifolds using mixture model cnns,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 5425–5434. [Online]. Available: <https://doi.org/10.1109/CVPR.2017.576>
 - [32] M. Niepert, M. Ahmed, and K. Kutzkov, “Learning convolutional neural networks for graphs,” in *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2016, pp. 2014–2023. [Online]. Available: <http://proceedings.mlr.press/v48/niepert16.html>
 - [33] H. Gao, Z. Wang, and S. Ji, “Large-scale learnable graph convolutional networks,” in *KDD*, 2018, pp. 1416–1424.
 - [34] W. L. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 1024–1034. [Online]. Available: <http://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs>
 - [35] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D. Yeung, “Gaan: Gated attention networks for learning on large and spatiotemporal graphs,” in *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, 2018, pp. 339–349. [Online]. Available: <http://auai.org/uai2018/proceedings/papers/139.pdf>
 - [36] J. B. Lee, R. A. Rossi, and X. Kong, “Graph classification using structural attention,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, 2018, pp. 1666–1674. [Online]. Available: <https://doi.org/10.1145/3219819.3219980>
 - [37] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding with unsupervised learning,” in *Technical report, OpenAI*.
 - [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008. [Online]. Available: <http://papers.nips.cc/paper/7181-attention-is-all-you-need>
 - [39] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186. [Online]. Available: <https://doi.org/10.18653/v1/n19-1423>
 - [40] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019, pp. 5754–5764. [Online]. Available: <http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding>
 - [41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: <http://arxiv.org/abs/1907.11692>
 - [42] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and P. Soricut, “ALBERT: A lite BERT for self-supervised learning of language representations,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. [Online]. Available: <https://openreview.net/forum?id=H1eA7AEtvS>