Collaborative Knowledge Distillation for Heterogeneous Information Network Embedding

Can Wang

wcan@zju.edu.cn

College of Computer Science, Zhejiang University
ZJU-Bangsun Joint Research Center

Sheng Zhou*

zhousheng_zju@zju.edu.cn

School of Software Technology, Zhejiang University
Zhejiang Provincial Key Laboratory of Service Robot

Kang Yu, Defang Chen, Bolang Li

{yk1083,defchern,21821210}@zju.edu.cn

College of Computer Science, Zhejiang University

Yan Feng, Chun Chen

{fengyan,chenc}@zju.edu.cn

College of Computer Science, Zhejiang University

ABSTRACT

Learning low-dimensional representations for Heterogeneous Information Networks (HINs) has drawn increasing attention recently for its effectiveness in real-world applications. Compared with homogeneous information networks, HINs are characterized by meta-paths connecting different types of nodes with semantic meanings. Existing methods mainly follow the prototype of independently learning meta-path-based embeddings and integrating them into a unified embedding. However, meta-paths in a HIN are inherently correlated since they reflect different perspectives of the same object. If each meta-path is treated as an isolated semantic data resource and the correlations among them are disregarded, sub-optimality in the both the meta-path based embedding and final embedding will be resulted. To address this issue, we make the first attempt to explicitly model the correlation among meta-paths by proposing **C**ollaborative **K**nowledge **D**istillation for Heterogeneous Information Network Embedding (**CKD**). More specifically, we model the knowledge in each meta-path with two different granularities: regional knowledge and global knowledge. We learn the meta-path-based embeddings by collaboratively distill the knowledge from intra-meta-path and inter-meta-path simultaneously. Experiments conducted on six real-world HIN datasets demonstrates the effectiveness of the CKD method.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Heterogeneous Information Networks, Network Embedding, Knowledge Distillation

*Corresponding Author

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512209

ACM Reference Format:

Can Wang, Sheng Zhou, Kang Yu, Defang Chen, Bolang Li, and Yan Feng, Chun Chen. 2022. Collaborative Knowledge Distillation for Heterogeneous Information Network Embedding. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485447.3512209

1 INTRODUCTION

Consisting of multiple types of nodes or edges, Heterogeneous Information Networks (HINs) [34] have been a powerful tool for modeling complex interactions, including social networks, bibliographic networks and biological networks[3, 17]. With the growth in network size and structural complexity, capturing the structural proximity in HINs with low-dimensional node representations is critical for downstream tasks, which has drawn increasing attention in both academic and industry communities.

Compared with the homogeneous network with a single type of nodes and edges, HINs are characterized by the diversified connection patterns among nodes, which are usually represented by *meta-path* [34] as ordered sequences of node types and edge types and nodes are connected by multiple meta-paths with different semantic meanings. Taking the bibliographic network as an example, a bibliographic network (Figure 1) typically consists of three types of nodes: author (A), paper (P), and Venue(V). Two authors are connected by Author-Paper-Author (APA) and Author-Paper-Venue-Paper-Author (APVPA), which describe the co-author and co-venue relationships between authors. Most existing HIN embedding methods [18, 30, 39] generally learn embeddings from each meta-path independently and fuse them into a unified one as the final output.

Despite effectiveness, we argue that these methods are not sufficient to yield satisfactory results. Although each meta-path contains specific semantic meaning of the HIN, they are inherently correlated since they reflect different perspectives of the same object. For instance, two researchers co-authoring a paper (meta-path APA) tend to share similar research interests and are more likely to submit papers to the same venue (meta-path APVPA). The two meta-paths are obviously correlated and knowledge from one meta-path would benefit another in learning its embedding. However, the prototype in existing methods treats each meta-path as an isolated semantic

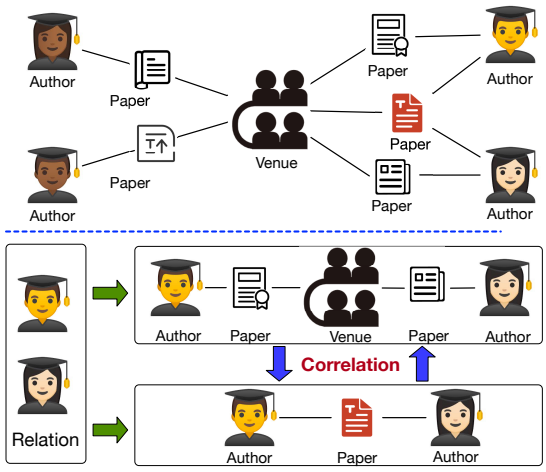


Figure 1: Heterogeneous information network and meta-path.

data resource and disregards the relationship among them in learning meta-path embeddings. Although meta-path embeddings will be eventually integrated,[9, 32, 47], the insufficiency in the intermediate results will lead to sub-optimality in the final representation.

Although important, it is challenging to simultaneously preserve both the semantic meaning for each meta-path and correlations among different meta-paths, due to their conflicting nature. A straightforward approach is to align meta-path embeddings by regularization. But this will homogenize the learned embedding for each meta-path, hurting the fundamental heterogeneity of the HIN. Inspired by the recent advances in collaborative knowledge distillation [1, 4, 45], we propose a novel **C**ollaborative **K**nowledge **D**istillation (CKD) framework to explicitly preserve the semantic meaning and the correlation among meta-paths. Our approach first adopts the graph diffusion and context sub-graph sampling strategy to address the varying sparsity issue for different meta-paths. We model the knowledge in each meta-path with two different granularities named regional knowledge and global knowledge. The knowledge in each meta-path are collaboratively distilled to enhance each other. Mutual information is used as the measure to guide the distillation in both intra-meta-path and inter-meta-path manner for learning a better embedding. Our major contributions are highlighted as follows:

(1) To the best of our knowledge, we make the first attempt to model the correlation between meta-paths in HIN embedding with the collaborative knowledge distillation framework.

(2) By modeling the regional and global knowledge in each meta-path, our approach can efficiently preserve local and global pattern in the final embedding by collaborative knowledge distillation in both inter-meta-path and intra-meta-path manner.

(3) Extensive experiments including node classification, link prediction, and ablation studies are conducted on six real-world HINs, which demonstrates the effectiveness of our proposed framework.

2 RELATED WORK

Network Representation Learning. Early Network Representation Learning methods learn node embedding by predicting the existance of edges [22, 48] or the proximity between nodes [35, 50]. Such methods may suffer from sparsity and scalability issues as many real-world networks are huge and sparse. Recently, Graph Neural Networks (GNNs) has been studied by aggregating information from neighborhood nodes, such as GCN [21], GraphSAGE [13], and GAT [37]. Recent models have also learned embedding in an unsupervised manner using mutual information and contrastive learning techniques[28, 38]. The methods mentioned above focus on networks with homogeneous nodes and edges which can not be naturally adapted to HINs. To learn node representations in HINs, Metapath2Vec preserves the proximity generated by random walks guided by meta-paths. HIN2Vec [11], HGT [18], and HAN [39] are representative Heterogeneous Graph Neural Networks (HGNNs) in which the heterogeneous networks are projected to homogeneous networks where GNNs are applied on. Recently, new efforts have been made on unsupervised HIN embedding. Among them, HDGI [30] utilizes the Infomax principles [16] on each meta-path based homogeneous graph, mg2vec [44] jointly embeds nodes and meta-graphs into the same space by exploiting both first-order and second-order proximity. Despite their success, existing methods have treated the meta-paths independently while the correlation among meta-paths can provide valuable insight into the semantic meanings.

Knowledge Distillation. Knowledge distillation was originally proposed by [15], where a complex and powerful model is called as a *teacher* model while a lightweight and relatively weaker model is called as a *student* model. To further boost the student performance, various approaches have been studied to align intermediate feature maps [5, 31, 42] or feature representations from the penultimate layer [36, 49]. To get rid of pretraining a large teacher model, online knowledge distillation is proposed by simultaneously training a group of student models, where each student model is encouraged to distill the knowledge from other peers [1, 4, 45]. Typically, they dynamically construct a target with higher accuracy to guide the training of each student, such as simply averaging the predictions from other peers [1, 45] or adaptively assigning the weights through self-attention mechanism [4]. To the best of our knowledge, we are the first attempt to collaboratively integrate the semantic knowledge from different meta-paths in HIN embedding with the idea of knowledge distillation.

Mutual Information Maximization. Mutual information measures the dependence among random variables. The conventional estimation of mutual information incurs expensive computational cost and can not apply on the high dimensional data like images. With the success of Mutual Information Neural Estimation (MINE) [2], mutual information can be efficiently estimated by training a statistics network as a classifier to distinguish samples coming from the joint distribution and the product of marginals of two random variables. Inspired by MINE, many mutual information estimators [27, 33] have been proposed and successfully applied on the unsupervised representation learning.

3 PRELIMINARIES

In this section, we give some formal definitions and important notations related to heterogeneous information network embedding.

DEFINITION 1. ***Heterogeneous Information Network (HIN)*** *is a type of information network whose vertices or edges are of different types. A heterogeneous informationn network can be represented as* $\mathcal{G} = \{\mathcal{V}, \mathcal{R}, \mathcal{E}\}$ *where* $\mathcal{V}$ *is the set of typed nodes,* $\mathcal{R}$ *is the set of edge types and* $\mathcal{E}$ *is the set of typed edges.*

DEFINITION 2. ***Meta-path*** *is a sequence of compatible edge types* $m = [r_1, r_2, ..., r_L]$ *defined in heterogeneous information network where* $r_l \in \mathcal{R}$ *is a specific type of edge.* ***Meta-path instance*** *is a sequence of nodes* $[v_1, v_2, ..., v_L]$ *that follows the connection order of the meta-path m.*

DEFINITION 3. ***Semantic Space*** *is defined as the homogeneous information network with single type of nodes, which is extracted by projecting the HIN* $\mathcal{G} = \{\mathcal{V}, \mathcal{R}, \mathcal{E}\}$ *with meta-path m.*

DEFINITION 4. ***Heterogeneous Information Network Embedding*** *aims at learning low-dimensional vector representation* $h_i \in \mathbb{R}^d$ *for each node* $v_i \in \mathcal{V}$ *in heterogeneous information network* $\mathcal{G}$ *so that the proximity between nodes can be preserved in the embedding space, d is the dimension of representation.*

4 MODEL

In this section, we propose a novel **C**ollaborative **K**nowledge **D**istillation (CKD) method for heterogeneous information network embedding. The CKD method consists three major components: semantic context subgraph sampling, heterogeneous knowledge modelling and collaborative knowledge distillation.

4.1 Semantic Context Subgraph Sampling

The success of Graph Neural Networks (GNNs) has proved the effectiveness of learning node embedding by aggregating information from neighborhood nodes. However, simply applying graph neural networks on different meta-paths for embedding learning is insufficient for HINs due to the following reasons:

(1) **Sparsity problem**. In some meta-paths, the relationship among nodes may be *valuable but sparse*. For example, the co-author relationship is critical for author classification while most authors may only have a limited number of co-authors. The graph neural networks applied on the meta-path APA can not aggregate enough information from limited neighborhoods.

(2) **Redundancy problem**. In some meta-paths, the relationship among nodes may be *abounding but redundant*. For example, each author can be connected to thousands of other authors by the co-venue relationship, while only a few of them are related. The graph neural networks applied on the meta-path APVPA will aggregate noisy information from redundant neighborhood.

Figure 2 illustrates the dataset analysis on three real-world HINs. The X axis denotes the selected meta-paths in the datasets and the Y axis denotes the average node degree in the homogeneous information network corresponding to the meta-path. We can observe that different meta-path based homogeneous information network has huge variation in density from same HIN. In the ACM dataset, the gap is up to 300 times between the node degree of Paper-Term-Paper(PTP) and Paper-Author-Paper(PAP).

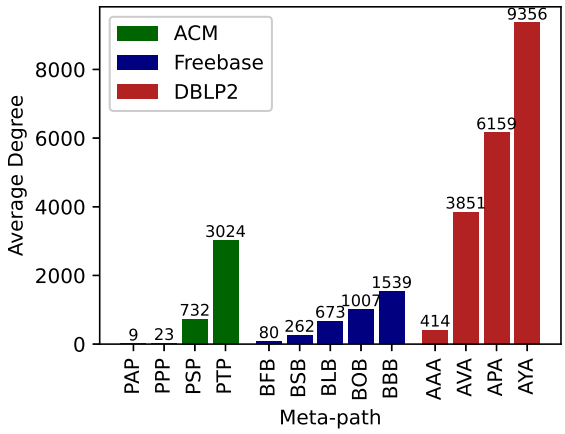


Figure 2: Data analysis on real-world HINs. The x-axis denotes the meta-path and y-axis denotes the average degree of nodes in the semantic space corresponding to the meta-path.

To this end, we utilize graph diffusion [23] technique to smooth out the neighborhood over the graph corresponding to different meta-paths. Then we sample a fixed size subgraph that contains sufficient structure information for meta-path based embedding learning. Given an HIN $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{R}\}$ and meta-path set $\mathcal{M}$, for each meta-path $m \in \mathcal{M}$, we first project the HIN into meta-path based homogeneous network G^m where $G^m_{ij} = 1$ if node v_i and v_j are connected by meta-path m, otherwise $G^m_{ij} = 0$. Then we utilize a generalized graph diffusion named Personalized PageRank (PPR) [10] to measure the connectivity between nodes in G^m as follows:

$$\mathbf{S}^m = \alpha \left(\mathbf{I}_n - (1-\alpha)\mathbf{D}_m^{-1/2}\mathbf{A}^m\mathbf{D}_m^{-1/2}\right)^{-1} \tag{1}$$

where $\mathbf{S}^m \in \mathbb{R}^{N\times N}$ is the diffusion matrix, N is the number of target type nodes, $\mathbf{A}^m \in \mathbb{R}^{N\times N}$ is the adjacent matrix of meta-path based homogeneous network G^m, $\mathbf{D}_m$ is the diagonal matrix with $\mathbf{D}_m(i,i) = \sum_j \mathbf{A}^m(i,j)$, and α is a parameter which is always set as 0.15. The diffusion matrix $\mathbf{S}^m$ has been proved effective in recovering meaningful neighborhoods from noisy graphs [23] which is essential for overcoming the varying sparsity and redundant problem of meta-path based homogeneous networks.

Intuitively, nodes are correlated to their regional neighbors while the distant provide limited information for node embedding learning. To further improve the efficiency, we propose a *semantic context subgraph sampling* strategy to sample the top-K important neighbors (including node itself) based on graph diffusion $\mathbf{S}^m$ to constitute a context subgraph C^m_i for meta-path based node embedding learning, which can be denoted as:

$$C^m_i = \text{top\_rank}(\mathbf{S}^m(i,:), K) \tag{2}$$

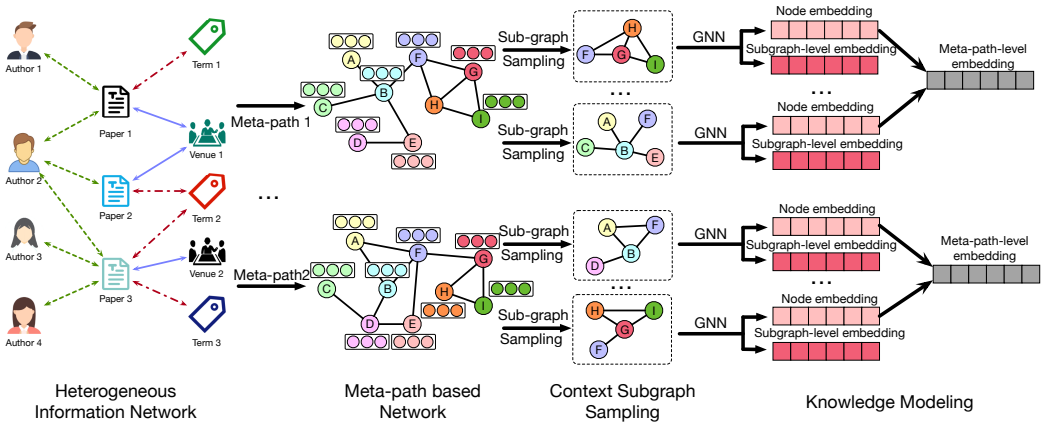


Figure 3: The forward propagation of CKD model and the definition of node embedding, subgraph-level embedding and meta-path-level embedding.

where K is the size of context graph and top_rank function returns the index of nodes ranked by the similarity. The diffusion matrix $\mathbf{S}^m$ can be precomputed before the model training starts and supports parallel computing so that the model can be applied to graphs that do not fit into GPU memory. The semantic context subgraph sampling is easy to parallel and scale to large datasets.

4.2 Heterogeneous Knowledge Modeling

Given the meta-path m defined on the HIN $\mathcal{G} = \{\mathcal{V}, \mathcal{R}, \mathcal{E}\}$, the HIN can be projected into a homogeneous information network $\mathcal{G}^m$ with semantic meanings. To explicitly distill knowledge for heterogeneous information network embedding, we first model two granularity of heterogeneous knowledge in each semantic space namely regional knowledge and global knowledge.

Given the meta-path based context subgraph C_i^m centered by node v_i, the node embedding h_i is learned by aggregating information from the context nodes in subgraph C_i^m with Graph Convolutional Network (GCN) [21]:

$$\mathbf{H}^m = \left(\widetilde{\mathbf{D}}_m^{-\frac{1}{2}} \tilde{\mathbf{A}}^m \widetilde{\mathbf{D}}_m^{-\frac{1}{2}}\right) \mathbf{X}^m \mathbf{W}^m \tag{3}$$

where $\mathbf{H}^m \in \mathcal{R}^{N \times d}$ is the d-dimensional meta-path based node embedding, $\tilde{\mathbf{A}}^m = \mathbf{A}^m + \mathbf{I}$ is the adjacent matrix of the context subgraph G^m with added self-connections, $\widetilde{\mathbf{D}}_m$ is the diagonal matrix corresponding to $\tilde{\mathbf{A}}^m$, $\mathbf{X}^m$ is the feature matrix, $\mathbf{W}^m$ is the trainable weight matrix for meta-path m. Note that other GNNs such as Graph Attention Networks [37] can also be applied here.

In each context subgraph C_i^m, we learn node embedding for both the centered node v_i and context node $v_{j,j\neq i}$. Each node $v_j \in \mathcal{V}$ can be context node of different center nodes in their context subgraphs with different local structures and context embeddings. We only use the centered node embedding h_i^m learned from subgraph C_i^m for optimization and downstream tasks.

4.2.1 Regional Knowledge. For each node $v_i \in \mathcal{V}$ in the HIN $\mathcal{G}$, they have personalized connection patterns with neighborhood nodes which can be reflected by the regional nodes around it. However, in the semantic space corresponding to different meta-paths, such connection pattern may vary which is not sufficient for the meta-path based embedding learning. To address this issue, we model the regional knowledge in each semantic space as the subgraph-level embedding around it.

Given the sampled context subgraph G_i^m centered around node v_i with meta-path m, subgraph-level embedding l_i^m, which is obtained by leveraging a local readout function $\mathcal{R}_l : \mathbb{R}^{(K)\times d} \rightarrow \mathbb{R}^d$ over the node embedding learned in each subgraph:

$$l_i^m = \mathcal{R}_l\left(G_i^m\right) = \sigma\left(\frac{1}{K}\sum_{j=1}^{K} h_j^m\right) \tag{4}$$

where h_j^m is the context embedding of nodes in context subgraph G_i^m including both centered node v_i and other context nodes, $\sigma(x) = 1/(1 + exp(-x))$ is the sigmoid function. Based on the above definition, a good meta-path based node embedding h_i^m is expected to distill the regional knowledge from both meta-path m and other meta-path $m' \in \mathcal{M}, m' \neq m$ so that the connection pattern can be well preserved.

4.2.2 Global Knowledge. Although the regional knowledge provides the local connection pattern of each node, the global connection pattern of the HIN that shared in all locations can not be preserved. To address this issue, we define the global knowledge as the *meta-path-level embedding* p^m, which is obtained by leveraging a global readout function $\mathcal{R}_g : \mathbb{R}^{N\times d} \rightarrow \mathbb{R}^d$ on the centered node embedding:

$$p^m = \mathcal{R}_g\left(H^m\right) = \sigma\left(\frac{1}{N}\sum_{i=1}^{N} h_i^m\right) \tag{5}$$

where h_j^m is the centered node-level embedding, σ is the nonlinear activation function. A good meta-path based node embedding is expected to distill the global pattern from all meta-paths. Figure 3 illustrates the forward propagation of our CKD model and the definition of node embedding, subgraph-level embedding and meta-path-level embedding.

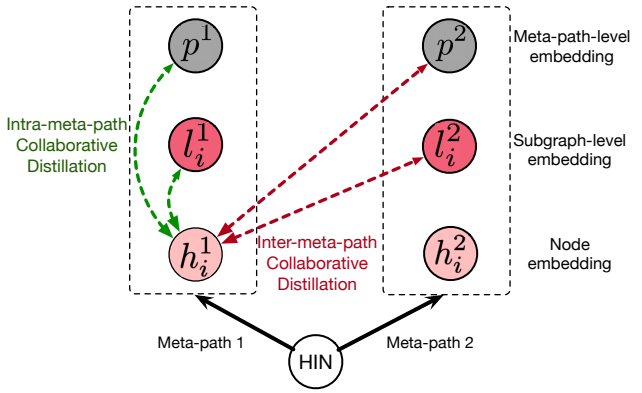


Figure 4: Collaborative Distillation Framework.

4.3 Collaborative Knowledge Distillation

Given the heterogeneous knowledge modelled above, a good meta-path based embedding is expected to distill regional knowledge and global knowledge from all the meta-path $m \in \mathcal{M}$, which is a typical *collaborative knowledge distillation* problem. We utilize Mutual Information (MI) as the measurement for distillation which has been widely used to capture non-linear statistical dependencies between variables.

4.3.1 Intra-meta-path Collaborative Distillation. The intra-meta-path collaborative distillation aims at simultaneously distill the regional and global knowledge within each meta-path to improve the node embedding. The distillation is measured by the mutual information between the node embedding h_i^m and subgraph-level embedding l_i^m, meta-path-level embedding p^m. The objective of intra-meta-path collaborative distillation is defined as:

$$\mathcal{L}_{intra} = -\sum_{m\in\mathcal{M}}\left(\sum_{i}^{|N|}\left(\mathrm{MI}(h_i^m, l_i^m) + \mathrm{MI}(h_i^m, p^m)\right)\right) \tag{6}$$

where MI is the mutual information estimator which will be introduced later.

4.3.2 Inter-meta-path Collaborative Distillation. The inter-meta-path collaborative distillation aims at simultaneously distill the regional and global pattern among different meta-paths. We use the similar way to distill as intra-meta-path collaborative distillation, while the difference is that embeddings are from different meta-paths. The objective of inter-meta-path collaborative distillation is defined as:

$$\mathcal{L}_{inter} = -\sum_{i}^{|N|}\left(\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{M},n\neq m}\mathrm{MI}(h_i^m, l_i^n) + \mathrm{MI}(h_i^m, p^n)\right) \tag{7}$$

4.3.3 Mutual Information Estimation. There are several mutual information estimators available including Jensen-Shannon divergence (JSD) and InfoNCE [16]. For simplicity, we utilize JSD-based mutual information estimator which is formulated as:

$$\mathrm{MI}(X, Y) = \mathbb{E}_{\mathbb{P}}[-sp(-f(x, y))] - \mathbb{E}_{\mathbb{P}\times\widetilde{\mathbb{P}}}[sp(f(x, \tilde{y}))] \tag{8}$$

where $sp(x) = log(1 + e^x)$ is the softplus function, $\mathbb{P}$ is the distribution of positive samples while $\widetilde{\mathbb{P}}$ is the distribution of negative samples.

For regional knowledge, the context subgraph embedding l_i^m from all meta-paths $l_i^n, n \in \mathcal{M}$ are regarded as positive samples, the other randomly sampled nodes are regarded as negative samples. For global pattern dependence, the meta-path-level embedding g^m can be regarded as positive sample, we follow the strategy of existing methods that first corrupt the node attributes and then learn meta-path-level embedding for the corrupted network as negative samples.

4.4 Model Training

The overall objective of our proposed CKD model is a combination of intra-meta-path and inter-meta-path collaborative knowledge distillation, which is formulated as:

$$\mathcal{L} = \mathcal{L}_{intra} + \mathcal{L}_{inter} \tag{9}$$

When the objective is optimized, we get the meta-path based node embedding h_i^m. The unified embedding is learned by simply summing up all the meta-path based node embedding:

$$h_i = \sum_{m\in\mathcal{M}} h_i^m \tag{10}$$

We adopt the Adam optimizer to minimize the objective and learn the optimal node embedding in an unsupervised manner.

Algorithm 1 CKD framework

Input: Heterogeneous Information Network $\mathcal{G} = \{\mathcal{V}, \mathcal{R}, \mathcal{E}\}$, the number of sampled neighborhood K, embedding dimension d
Output: Node, Subgraph-level and Meta-path-level Embedding h_i^m, l_i^m, p^m

```
Sample semantic context subgraph C_i^m for each node v_i ∈ V with graph diffusion.
while Not Converged do
    Forward propagation and define regional and global knowledge defined in Eq 3,4,5.
    Perform intra-meta-path and inter-meta-path collaborative knowledge distillation with Eq 6,7.
    Back propagate and update three embedding with Eq 9,10.
end while
Return three embedding h_i^m, l_i^m, p^m.
```

5 EXPERIMENTS

In this section, we conduct extensive experiments on six real-world heterogeneous information networks with respect to node classification, link prediction and ablation studys. The experiments are designed to answer the following research questions:

RQ1 Does CKD outperform state-of-the-art methods?

RQ2 Is it beneficial to introduce the collaborative knowledge distillation for HIN embedding?

RQ3 How does the parameter affect the performance?

5.1 Datasets

We conduct experiments on six real-world HIN datasets named Pubmed, DBLP, ACM and Freebase[41], which have been widely used by existing methods. The brief statistics of the datasets are illustrated in Table 1.

Pubmed. The Pubmed network consists of four types of nodes, including genes(G), diseases(D), chemicals(C), and species(S). The links represent the semantic relationships among objects. Each node in the network is associated with features concerning the words in the paper. The disease nodes are labeled into 8 unique categories.

DBLP/DBLP2. The DBLP network is a bibliographic network where nodes represent authors(A), papers(P), venues(V), and phrases(P). The papers and phrase features are computed by word2vec [26] on all paper texts. The author and venue features are the aggregations of the corresponding paper features. The author nodes are labeled into 4 unique research groups from four research areas.

ACM/ACM2. The ACM network is another bibliographic network where nodes represent papers(P) and authors(A). The paper notes are labeled into 3 basic unique classes, including database, wireless communication, data Mining or 7 smaller classes. The features of nodes are the Word2Vec [25] based embedding of abstract.

Freebase Freebase was a large collaborative knowledge network consisting of books(BO), films(F), music(M), sports(S), people(P), locations(L), organizations(O) and businesses(BU). Since nodes in the original dataset are not associated with features, we follow the experimental setting of the existing method [46] and apply DeepWalk [29] to generate node feature. The books are labeled into 7 genres of literature.

Table 1: Statistics of the datasets used in the experiments.

Dataset	Nodes	Edges	Features	Labels
ACM	10,942	547,872	100	3
ACM2	29,930	61,770	100	7
DBLP	26,128	239,566	200	4
DBLP2	173,988	20,743,972	300	4
Pubmed	63,109	125,167	200	8
Freebase	79,843	498,508	300	7

5.2 Baselines

We compare CKD framework with one classic homogeneous information network embedding method and nine state-of-the-art HIN embedding methods:

DeepWalk [29] performs a random walk on homogeneous network and then learns the representation of nodes via the Skip-Gram model. We compare with this method since it is used to generate node features for some datasets without node attributes.

Metapath2Vec [8] uses meta-path guided random walk to generate heterogeneous node sequences with rich semantcis. The heterogenous skip-gram technique is utilized to preserve the proximity between nodes.

HIN2Vec [11] carries out multiple relation prediction tasks jointly to learn the embeddings of nodes and meta-paths in heterogeneous informaiton networks.

HAN [39] utilizes a hierarchical attention mechanism to capture both node and semantic importances of meta-paths.

HDGI [30] utilizes the attention mechanism to capture the semantice meaning of meta-paths and preserve the diverse patterns in heterogeneous informaiton networks by maximizing the local-global mutual information.

HGT [18] utilizes attention mechanism to calculate different importance of neighborhood nodes around target nodes and assign weights during aggregation.

NSHE [46] learns embeddings by preserving pairwise structure and network schema structure concurrently, the heterogeneity within HINs is preserved by multi-tasks classifiers.

MAGNN [12] utilizes three components to learn node embedding named node content transformation, intra-meta-path aggregation and inter-meta-path aggregation. However, the aggregation is performed on embedding level which can not model the dependence.

HetGNN [43] preserves the heterogeneity of both graph structures and node attributes by aggregating node content, neighborhood information and meta-paths.

HeCo [40] is a co-contrastive learning method for heterogeneous graph neural networks, which performs contrastive learning on network schema and meta-path views.

For the baseline methods used in our experiment, we use the authors' open-sourced codes. For models that use meta-paths in modeling, we choose the popular meta-paths adopted in previous methods and report the best results. The code and dataset used in our experiments has been published in Github [1].

5.3 Node Classification

In this subsection, we evaluate the performance of node embedding with node classification tasks. We follow the experimental setting of existing unsupervised HIN embedding methods[7, 14, 19] and first learn the node embedding for target type nodes in an unsupervised manner. After having obtained the node representations, we randomly sample particular percentage (1/3,1/4,1/5) of labeled nodes to train a SVM classifier, and the rest of the nodes are used to test performances. We report the average performances in terms of both Macro-F1 and Micro-F1 score.The detailed results are shown in Table 2. To summarize, we have the following observations:

(1) Compared with the baseline methods, our proposed CKD framework achieves better node classification performance in most evaluated datasets. This proves the effectiveness of collaborative knowledge distillation for heterogeneous information network embedding.

(2) In some large scale datasets such as Freebase and DBLP2, methods like NSHE, MAGNN, HeCo and HetGNN can not scale to such datasets. However, the porposed CKD framework can scale to such datasets and achieves comparable performance. This demonstrate

[1] https://github.com/zhoushengisnoob/CKD

Table 2: Node classification on six real-world HINs. Bold fonts denote the best performance among all methods. '-' denotes that the method can not be run on our hardware settings. Each method has three lines corresponding to 1/3,1/4,1/5 data for training classifier.

Dataset	ACM		DBLP		ACM2		PubMed		Freebase		DBLP2	
Metric	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
DeepWalk	89.6±0.0	89.6±0.0	91.3±0.0	91.7±0.0	64.8±0.0	75.9±0.0	15.1±0.0	16.8±0.0	**48.9**±0.0	**60.6**±0.0	88.4±0.0	88.3±0.0
	88.8±0.0	88.8±0.0	90.6±0.0	91.0±0.0	64.8±0.0	75.9±0.0	14.7±0.0	16.5±0.0	48.1±0.0	60.1±0.0	88.4±0.0	88.2±0.0
	89.8±0.0	89.8±0.0	90.8±0.0	91.2±0.0	64.6±0.0	76.0±0.0	12.9±0.0	15.7±0.0	**49.3**±0.0	**60.8**±0.0	88.3±0.0	88.1±0.0
Metapath2Vec	91.3±0.3	91.4±0.3	86.3±1.0	87.0±0.9	38.3±1.2	59.0±1.4	13.7±1.2	15.5±1.0	42.2±0.4	54.7±0.2	87.8±0.3	87.6±0.3
	91.7±0.6	91.8±0.6	87.7±1.0	88.3±0.9	38.9±1.1	59.1±1.3	12.4±1.4	14.5±1.2	41.5±1.0	54.6±0.3	88.0±0.2	87.8±0.3
	92.0±0.5	92.1±0.5	89.2±0.5	89.4±0.8	38.8±1.1	59.3±1.5	13.2±1.1	15.2±1.1	41.6±0.3	54.9±0.3	87.9±0.2	87.8±0.1
HIN2Vec	88.5±1.2	88.4±1.3	92.2±0.2	92.5±0.3	23.4±0.6	53.9±0.3	14.8±0.7	18.4±0.5	26.4±0.7	49.3±0.9	86.9±0.4	86.7±0.5
	89.6±1.8	89.8±1.7	91.9±0.2	92.4±0.2	23.7±0.5	54.9±0.6	14.2±0.5	17.8±0.3	25.9±0.4	49.5±0.7	86.6±0.4	86.8±0.3
	89.8±1.6	89.7±1.8	92.5±0.3	93.0±0.2	26.8±0.7	57.4±0.5	14.5±0.8	17.6±0.6	26.0±0.5	49.5±0.8	87.5±0.2	87.3±0.3
HAN	90.4±1.2	90.5±1.2	88.0±0.5	88.5±0.5	59.2±0.9	74.5±0.6	35.1±0.5	37.5±0.3	46.5±0.5	60.1±0.6	88.1±0.6	88.1±0.6
	90.7±1.4	90.8±1.3	87.6±0.7	88.1±0.4	58.7±1.1	74.0±0.8	34.3±0.7	37.1±0.5	46.6±1.1	60.9±0.6	87.5±1.3	87.4±1.4
	90.5±1.0	90.5±1.0	88.4±0.8	88.9±0.8	59.1±0.8	74.5±0.6	35.0±0.8	38.5±0.6	46.7±0.8	60.9±0.4	88.2±0.7	88.2±0.7
HDGI	68.8±2.4	68.9±2.1	74.4±1.0	75.9±1.0	31.5±1.2	57.1±1.2	14.9±0.8	20.3±0.6	-	-	86.8±0.8	87.0±0.8
	68.8±2.2	68.4±2.1	74.5±1.2	75.9±1.1	31.7±1.3	57.2±1.2	15.2±0.7	20.5±0.5	-	-	87.0±0.9	87.2±0.8
	69.8±2.7	69.5±2.8	74.5±1.3	76.0±1.4	31.8±1.4	57.4±1.2	15.4±0.6	20.7±0.4	-	-	87.1±0.9	87.2±0.8
HGT	89.1±0.4	89.3±0.3	50.8±1.0	50.7±1.2	60.9±1.0	75.4±1.2	19.0±0.5	19.9±0.8	-	-	84.1±0.6	84.3±0.6
	89.1±0.5	89.3±0.4	50.9±1.2	51.0±1.1	61.1±1.1	75.7±1.3	20.6±1.9	22.0±1.3	-	-	84.2±0.6	84.4±0.7
	89.2±0.7	89.3±0.7	52.7±0.7	52.8±0.6	61.3±1.2	75.8±1.3	19.4±2.5	20.7±3.7	-	-	84.3±0.9	89.2±0.9
NSHE	90.3±0.3	90.4±0.2	**93.9**±0.1	**94.1**±0.2	62.4±0.6	75.9±0.2	17.1±0.7	22.3±0.9	-	-	-	-
	90.5±0.2	90.6±0.2	**93.8**±0.3	**94.0**±0.3	62.4±0.7	75.9±0.2	17.5±0.8	22.7±0.6	-	-	-	-
	89.7±0.3	89.8±0.3	**93.9**±0.2	**94.1**±0.2	62.5±0.8	76.1±0.2	17.7±0.8	22.9±1.1	-	-	-	-
MAGNN	85.7±0.2	85.7±0.2	87.9±0.3	88.3±0.4	51.0±0.8	70.8±0.4	34.1±1.2	38.3±0.9	47.1±0.6	60.1±0.3	-	-
	87.3±0.4	87.3±0.4	87.5±0.5	88.3±0.2	52.1±0.7	67.8±1.1	36.3±0.6	38.9±0.7	47.6±0.3	60.0±0.5	-	-
	87.9±0.4	88.0±0.4	88.2±0.8	88.9±0.5	53.8±0.6	70.8±0.7	**39.4**±0.7	**42.1**±0.8	47.4±0.7	60.4±0.4	-	-
HeCo	71.0±0.2	71.2±0.1	91.5±0.5	91.8±0.6	57.2±0.8	72.9.0±0.5	16.5±0.5	26.1±1.2	-	-	-	-
	71.2±0.4	71.3±0.3	91.2±0.5	91.4±0.6	56.7±0.9	73.0±0.3	16.8±0.6	25.7±1.1	-	-	-	-
	71.3±0.1	71.3±0.1	91.2±0.4	91.5±0.5	57.5±1.1	72.9±0.7	16.9±0.7	25.9±1.0	-	-	-	-
HetGNN	85.7±0.1	85.6±0.1	92.0±0.6	92.3±0.7	-	-	-	-	-	-	-	-
	86.1±0.1	86.1±0.1	92.3±0.5	92.6±0.5	-	-	-	-	-	-	-	-
	86.6±0.2	86.7±0.2	92.8±0.6	93.1±0.5	-	-	-	-	-	-	-	-
CKD	**91.9**±0.4	**91.9**±0.4	92.5±0.2	92.8±0.2	**69.7**±0.5	**79.7**±0.8	**36.8**±1.1	**39.3**±1.6	48.2±0.7	60.5±0.4	**90.2**±0.3	**90.1**±0.3
	92.9±0.3	**92.9**±0.3	92.5±0.4	92.8±0.4	**65.6**±0.3	**77.9**±0.1	**37.4**±0.9	**40.1**±0.6	**49.6**±0.4	**61.1**±0.7	**90.4**±0.3	**90.3**±0.3
	92.8±0.8	**92.7**±1.0	92.3±0.4	92.6±0.4	**70.4**±0.5	**80.2**±0.6	37.8±1.2	40.4±1.2	48.1±0.8	60.4±0.5	**90.2**±0.2	**90.1**±0.1

the advantage of context subgraph sampling designed in the CKD framework.

(3) An interesting observation is that DeepWalk achieves good performance in the Freebase dataset. This is explainable since we use embedding generated by DeepWalk as node attributes in Freebase. The GNN based methods suffer from aggregating redundant information from neighborhood nodes.

5.4 Link Prediction

We further evaluate node embedding performance with link prediction, which is another basic application of node embedding[19, 20]. Given a heterogeneous information network, we first generate a training network by selecting an edge class and randomly removing a certain fraction (20% in our experiments) of edges from the selected edge class as missing edges. After removing edges, we apply representation learning models to the resulting sub-network. Then, to perform the link prediction on the training network, we apply supervised models to rank node pairs that are more likely to have missing edges. Table 3 illustrates the results of link prediction, from which we have the following observations:

Table 3: Performance on the link prediction task on three real-world datasets. Bold fonts denote the best performance among all compared methods. '-' denotes that the method can not be run on our hardware settings.

Method \ Data	ACM2	DBLP2	PubMed
DeepWalk	0.818	0.789	0.663
Metapath2Vec	0.712	0.915	0.628
HIN2Vec	0.736	0.803	0.649
HAN	0.868	0.711	0.717
HDGI	0.537	0.691	0.594
HGT	0.920	0.868	0.736
NSHE	0.939	-	0.654
MAGNN	0.696	-	0.514
HeCo	0.681	-	0.519
CKD	**0.948**	**0.931**	**0.735**

(1) Similar to node classification task, the CKD framework achieves better performance than the existing methods in three real-world HIN datasets. This further demonstrates the effectiveness of CKD framework in modeling the topology structure in HINs.

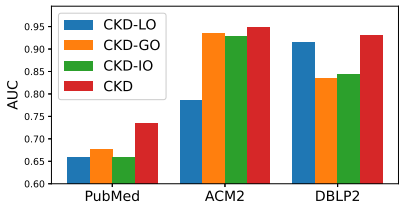


Figure 5: Ablation Study on the link prediction task on three real-world datasets.

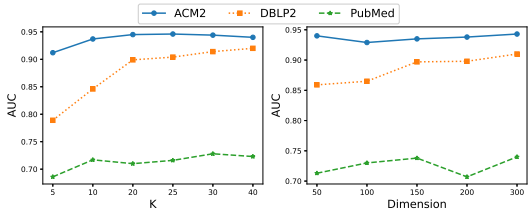


Figure 6: Parameter Analysis of the CKD framework with respect to the size of context subgraph and dimension of embedding.

(2) Compared with node classification task, the CKD and HGT method still achieve competitive performance. In contrast, methods like MAGNN and NSHE failed to do so. This indicates the over-reliance on structures suffer from the incomplete network structure.

5.5 Ablation Study

In order to verify the effectiveness of the delicate designs in the proposed CKD framework, we propose variants of CKD as follows:

- **CKD-LO** is the *regional only* variant of CKD which only perform collaborative distillation on the regional knowledge.
- **CKD-GO** is the *global only* variant of CKD which only perform collaborative distillation on the global knowledge.
- **CKD-IO** is the *intra-meta-path only* variant of CKD which only perform intra-meta-path collaborative knowledge distillation.

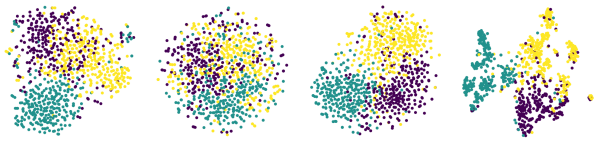
Figure 5 illustrates the ablation study on the variants of CKD framework on the link prediction task. We can have the following observations: The fact that CKD achieves better performance than CKD-LO, CKD-GO, CKD-IO indicates the necessity of distilling the regional knowledge and global knowledge in a collaborative distillation manner. Comparing CKD-LO and CKD-GO, we can find that different HIN datasets has different characteristics on the regional and global knowledge. This further confirms the effectiveness of distilling the regional and global knowledge in a unified framework. To conclude, the ablation study confirms the effectiveness of modules designed in CKD framework.

5.6 Parameter Analysis

In this section, we investigate the sensitivity of parameters and report the results of CKD in terms of link prediction on real-world datasets with different parameters.

Size of Context Subgraphs. In CKD model, a key parameter is the size of the context subgraph K. We vary the context subgraph size K from 5 to 50 and Figure 6-(a) illustrated the experimental results. We can observe that in smaller datasets like Freebase, ACM, and Pubmed, the performance is stable when K is larger than 20. In DBLP, the performance is stable after K is larger than 40. This is explainable since the DBLP dataset is dense. The small context subgraph can not sufficiently represent the local structural information. Since smaller K refers to fewer parameters and faster training, the CKD framework can achieve good performance with the lowest computational complexity with context subgraph sampling.

Embedding Dimension. We also evaluate the impact of the embedding dimensions on CKD task by varying the dimension between 25 and 125. As shown in Figure 6-(b), the performance of CKD model is generally robust when the dimensions are set to around 50, which is also a typical choice established by previous work on network embedding.



(a) HetGNN (b) MAGNN (c) HIN2Vec (d) CKD

Figure 7: Node Visualization results on ACM dataset.

5.7 Visualization

Following the experimental setting of existing works[6, 24], we also perform the network visualization experiments on the ACM dataset. We first learn a low dimensional representation for each node and then map those representations into the 2-D space with t-SNE. Figure 7 shows the results, from which we can observe the proposed CKD framework are quite clear since most of nodes with same label (color) are close to each other and nodes with different labels(colors) are far from each other. This further verifies the effectiveness of the proposed CKD method.

6 CONCLUSION

In this paper, we propose CKD framework, which is the first attempt of collaborative knowledge distillation for heterogeneous information network embedding. We first model the regional knowledge and global knowledge in each meta-path, then we collaboratively distill the knowledge from both the intra-meta-path and inter-meta-path manner. The mutual information is used as the measure to guide the distillation process. We conduct extensive experiments on six real-world HIN datasets and the results demonstrate the effectiveness of the proposed CKD framework.

7 ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (Grant No: 62106221, U1866602), Alibaba-Zhejiang University Joint Institute of Frontier Technologies and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (Grant No. SN-ZJU-SIAS-001).

REFERENCES

[1] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Róbert Ormándi, George E. Dahl, and Geoffrey E. Hinton. 2018. Large scale distributed neural network training through online distillation. In *International Conference on Learning Representations*.

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. 2018. Mine: mutual information neural estimation. *ArXiv preprint* (2018).

[3] Yukuo Cen, Xu Zou, Jianwei Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Representation learning for attributed multiplex heterogeneous network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1358–1368.

[4] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. 2020. Online Knowledge Distillation with Diverse Peers.. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3430–3437.

[5] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. 2021. Cross-Layer Distillation with Semantic Calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7028–7036.

[6] Hongxu Chen, Hongzhi Yin, Weiqing Wang, Hao Wang, Quoc Viet Hung Nguyen, and Xue Li. 2018. PME: projected metric embedding on heterogeneous networks for link prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1177–1186.

[7] Xia Chen, Guoxian Yu, Jun Wang, Carlotta Domeniconi, Zhao Li, and Xiangliang Zhang. 2019. Activehne: Active heterogeneous network embedding. *arXiv preprint arXiv:1905.05659* (2019).

[8] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *SIGKDD*. 135–144.

[9] Yuxiao Dong, Ziniu Hu, Kuansan Wang, Yizhou Sun, and Jie Tang. 2020. Heterogeneous Network Representation Learning.. In *IJCAI*, Vol. 20. 4861–4867.

[10] Dániel Fogaras, Balázs Rácz, Károly Csalogány, and Tamás Sarlós. 2005. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics* (2005), 333–358.

[11] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1797–1806.

[12] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding. In *The Web Conference*. 2331–2341.

[13] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeuraIPS*. 1024–1034.

[14] Yu He, Yangqiu Song, Jianxin Li, Cheng Ji, Jian Peng, and Hao Peng. 2019. Hetespaceywalk: A heterogeneous spacey random walk for heterogeneous information network embedding. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 639–648.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[16] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *ArXiv preprint* (2018).

[17] Huiting Hong, Hantao Guo, Yucheng Lin, Xiaoqing Yang, Zang Li, and Jieping Ye. 2020. An attention-based graph neural network for heterogeneous structural learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4132–4139.

[18] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *The Web Conference*. 2704–2710.

[19] Xiao Huang, Qingquan Song, Fan Yang, and Xia Hu. 2019. Large-scale heterogeneous feature embedding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3878–3885.

[20] Rana Hussein, Dingqi Yang, and Philippe Cudré-Mauroux. 2018. Are meta-paths necessary? Revisiting heterogeneous graph embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 437–446.

[21] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *ArXiv preprint* (2016).

[22] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *ArXiv preprint* (2016).

[23] Johannes Klicpera, Stefan Weißenberger, and Stephan Günnemann. 2019. Diffusion improves graph learning. In *NeuraIPS*. 13354–13366.

[24] Yuanfu Lu, Chuan Shi, Linmei Hu, and Zhiyuan Liu. 2019. Relation structure-aware heterogeneous information network embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4456–4463.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *NeuraIPS* (2013), 3111–3119.

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[28] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. 2020. Graph Representation Learning via Graphical Mutual Information Maximization. In *The Web Conference*. 259–270.

[29] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[30] Yuxiang Ren, Bo Liu, Chao Huang, Peng Dai, Liefeng Bo, and Jiawei Zhang. 2019. Heterogeneous deep graph infomax. *ArXiv preprint* (2019).

[31] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *Proceedings of the International Conference on Learning Representations*.

[32] Chuan Shi, Yuanfu Lu, Linmei Hu, Zhiyuan Liu, and Huadong Ma. 2020. RHINE: Relation structure-aware heterogeneous information network embedding. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[33] Jiaming Song and Stefano Ermon. 2019. Understanding the limitations of variational mutual information estimators. *ArXiv preprint* (2019).

[34] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB* (2011), 992–1003.

[35] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *The Web Conference*. 1067–1077.

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *Proceedings of the International Conference on Learning Representations*.

[37] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *ArXiv preprint* (2017).

[38] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep graph infomax. (2019).

[39] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The Conference*. 2022–2032.

[40] Xiao Wang, Nian Liu, Hui Han, and Chuan Shi. 2021. Self-supervised Heterogeneous Graph Neural Network with Co-contrastive Learning. *arXiv preprint arXiv:2105.09111* (2021).

[41] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. 2020. Heterogeneous Network Representation Learning: Survey, Benchmark, Evaluation, and Beyond. *TKDE* (2020).

[42] Sergey Zagoruyko and Nikos Komodakis. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the International Conference on Learning Representations*.

[43] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *SIGKDD*. 793–803.

[44] Wentao Zhang, Yuan Fang, Zemin Liu, Min Wu, and Xinming Zhang. 2020. mg2vec: Learning Relationship-Preserving Heterogeneous Graph Representations via Metagraph Embedding. *TKDE* (2020).

[45] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4320–4328.

[46] Jianan Zhao, Xiao Wang, Chuan Shi, Zekuan Liu, and Yanfang Ye. 2020. Network Schema Preserving Heterogeneous Information Network Embedding.

[47] Sheng Zhou, Jiajun Bu, Xin Wang, Jiawei Chen, and Can Wang. 2019. Hahe: Hierarchical attentive heterogeneous information network embedding. *arXiv preprint arXiv:1902.01475* (2019).

[48] Sheng Zhou, Xin Wang, Jiajun Bu, Martin Ester, Pinggang Yu, Jiawei Chen, Qihao Shi, and Can Wang. 2020. DGE: Deep Generative Network Embedding Based on Commonality and Individuality.. In *AAAI*.

[49] Sheng Zhou, Yucheng Wang, Defang Chen, Jiawei Chen, Xin Wang, Can Wang, and Jiajun Bu. 2021. Distilling Holistic Knowledge with Graph Neural Networks. In *International Conference on Computer Vision*.

[50] Sheng Zhou, Hongxia Yang, Xin Wang, Jiajun Bu, Martin Ester, Pinggang Yu, Jianwei Zhang, and Can Wang. 2018. Prre: Personalized relation ranking embedding for attributed networks. In *CIKM*. 823–832.