# Non-local Attention Learning on Large Heterogeneous Information Networks

Yuxin Xiao*, Zecheng Zhang*, Carl Yang, and Chengxiang Zhai
*Department of Computer Science, University of Illinois at Urbana-Champaign*
Urbana, Illinois, USA
{yuxinx2, zzhan147, jiyang3, czhai}@illinois.edu

*Abstract*—**Heterogeneous information network (HIN) summarizes rich structural information in real-world datasets and plays an important role in many big data applications. Recently, graph neural networks have been extended to the representation learning of HIN. One very recent advancement is the hierarchical attention mechanism which incorporates both node-wise and semantic-wise attention. However, since HIN is more likely to be densely connected given its diverse types of edges, repeatedly applying graph convolutional layers can make the node embeddings indistinguishable very quickly. In order to avoid oversmoothness, existing graph neural networks targeting HIN generally suffer from a shallow structure. Consequently, those approaches ignore information beyond the local neighborhood. This design flaw violates the concept of non-local learning, which emphasizes the importance of capturing long-range dependencies. To properly address this limitation, we propose a novel framework of non-local attention in heterogeneous information networks (NLAH). Our framework utilizes a non-local attention structure to complement the hierarchical attention mechanism. In this way, it leverages both local and non-local information simultaneously. Moreover, a weighted sampling schema is designed for NLAH to reduce the computation cost for large-scale datasets. Extensive experiments on three different real-world heterogeneous information networks illustrate that our framework exhibits extraordinary scalability and outperforms state-of-the-art baselines with significant margins.**

## I. Introduction

Graphs provide a natural way to represent many kinds of data and information in the real world, such as social networks, knowledge graphs, and the world wide web. Once represented as graphs in the non-Euclidean space, the data can be analyzed flexibly by using many powerful graph mining algorithms.

Heterogeneous information network (HIN) [1], [2] refers to graphs with multiple types of nodes and edges. In Figure 1, the DBLP dataset can be constructed into a heterogeneous information network. The network consists of three types of nodes: author (A), paper (P) and conference (C), and two kinds of edges: authoredBy (between paper and author) and publishedIn (between paper and conference). As illustrated by this example, many real-world relational datasets can be naturally summarized by the HIN framework, thus developing effective models, especially general models, for analyzing HIN has broad impacts in many big data application domains [1].

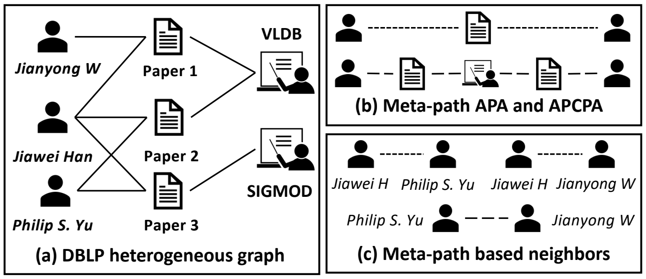


Fig. 1: An example of heterogeneous information network in DBLP dataset. (a) DBLP heterogeneous information network contains three types of nodes and two types of edges. (b) Two meta-paths APA and APCPA. (c) Meta-path based neighbors.

In order to better capture the rich semantic information in graphs, the graph representation learning algorithms embed graphs in lower-dimensional spaces and benefit various downstream tasks [3]–[7]. In particular, graph neural networks (GNNs) [8]–[10] introduce the neural convolution framework into the graph mining domain and evoke great research interests in recent years.

However, most existing GNN-based graph embedding methods can only deal with homogeneous networks and how to extend them to heterogeneous information networks has not been fully explored. The diverse types of nodes and edges in HIN pose a unique challenge for this task, where traditional GNN-based methods cannot be directly applied. Different types of nodes and edges in HIN typically carry different kinds of information, and hence, should be treated separately.

One of the latest heterogeneous information network embedding methods, HAN [11] attempts to mitigate this challenge by leveraging the attention mechanism to distinguish the importance of different neighbors during embedding aggregation, but its hierarchical attention mechanism has several shortcomings. (1) **Lack of non-local learning:** Since GNN is a special form of Laplacian smoothing [12], stacking multiple GNN layers may oversmooth features of nodes from different clusters and reduce the discriminative power of graph embedding. This phenomenon is even more concerning when GNN is applied to the embedding of HIN. Since HIN is more likely to be densely connected given its diverse types of edges, Laplacian smoothing [13] can happen even more quickly. As a result, GNN-based models for HIN embedding typically have a

*Equal contribution.

978-1-7281-0858-2/19/$31.00 © 2019 IEEE

shallow structure, which limits the models' effective receptive fields to the local neighborhood. On the other hand, as argued in [14], capturing long-range dependencies is the key to deep neural networks. Therefore, the study of non-local learning in HIN embedding is a pressing issue. (2) **Scalability concern:** Another limitation of the hierarchical attention mechanism is that the calculation of the node-wise attention involves all the neighbor pairs. This is computationally infeasible when the mechanism is applied to large dense networks.

In this paper, we address both limitations of HAN and propose a novel framework of non-local attention in heterogeneous networks (NLAH). Instead of adding more neural layers, we design a non-local attention structure to capture long-range latent relationships. More specifically, NLAH first computes non-local features with respect to each target node using properly designed non-local measures. It then creates a virtual neighbor for the target node to represent the generated non-local features. This virtual neighbor is fed into the hierarchical attention mechanism together with other real neighbors of the target node. In this way, the introduction of non-local features to the local neighborhood allows nodes to attend to both local and non-local information at the same time. In addition to that, we also investigate the requirements for proper non-local measures, and subsequently present three different non-local measures that significantly boost the performance of our framework. To address the scalability concern, we carry out node-wise sampling in NLAH based on the relative amount of information contained in each edge and each meta-path based sub-graph. This action reduces the training variance and computation cost.

To summarize, we make the following contributions:

- To our best knowledge, this is the first attempt to study the concept of non-local learning in HIN embedding. We propose a novel framework of non-local attention in heterogeneous information networks (NLAH) which leverages a non-local attention structure to complement the hierarchical attention mechanism.
- We conduct node-wise sampling according to the information density of each edge and each meta-path based sub-graph. Besides, our model can be efficiently implemented via parallelization. This further reduces the computation cost when the model is applied to real-world large-scale datasets.
- We examine the performance of our framework based on three different types of real-world heterogeneous information networks. The extensive experimental results demonstrate the superior effectiveness and efficiency of our model through comparison with various state-of-the-art baselines.

## II. Related Work

Our framework draws inspiration from the area of heterogeneous information network, the concept of non-local learning and the recent advancements in applying neural network to semi-supervised learning over graphs. In what follows, we would like to give a brief overview of these fields.

**Graph Neural Network.** Graph-based semi-supervised learning has been a popular research area for decades. [9], [15], [16] utilize the spectral graph convolutions to aggregate local structural features and node attributes. GAT [10] introduces the attention mechanism into feature aggregation by implicitly learning different weights to different node neighbors. Graph Attention Model [17], instead, leverages the attention mechanism on the power series of the transition matrix to optimize an upstream objective. Jumping Knowledge Network [18] proposes to aggregate embedding features across stacked neural layers to enable better structure-aware representation. GraphSAGE [8] implements node-wise sampling to get a fixed number of neighbors for node embedding. Self-paced learning [19] samples negative context nodes in terms of their informativeness. FastGCN [20] directly samples nodes via importance sampling. Our framework can be regarded as an extension of GNN to the field of HIN which poses special challenges for representation learning.

**Non-local Learning.** This concept emphasizes the inclusion of non-local information to capture long-range dependencies. Non-local Neural Network [14] proposes a non-local block in the neural network architecture in computer vision domain. Some graph neural network frameworks also consider non-local influence in their graph embedding tasks. LINE [21] employs the second-order proximity to preserve global structure of graphs. [12] relieves GCN's problem of localization with the use of self-training and co-training. DGCN [22] regularizes local consistency with global consistency during the training process of GCN. APPNP [23] propagates neural network predictions by adopting a personalized PageRank scheme. Different from these previous works, we focus on applying the concept of non-local learning to the new task of HIN embedding.

**Heterogeneous Information Network.** Heterogeneous information network (HIN) contains different types of nodes and edges, which better reflects the real scenario. PathSim [24] raises the concept of meta-path that defines different semantic meaning through sequences of relations. ESim [25] incorporates users' guidance on multiple meta-paths in their embedding framework. Metapath2vec [26] is a random walk based approach and utilizes skip-gram to perform heterogeneous information network embedding. HIN2Vec [27] is designed to capture the rich semantics embedded in HINs by exploiting different types of relationships among nodes. GraphInception [28] focuses on the extraction of a hierarchy of relational features in HIN by introducing the Graph Inception module. HAN [11] proposes a hierarchical attention mechanism, which consists of two stages: node-wise attention and semantic-wise attention. Our work attempts to resolve HAN's problem of localization by applying the concept of non-local learning.

## III. Hierarchical Attention in HIN

HAN [11] is one of the latest advancements in the area of heterogeneous information network representation learning. As our framework is an extension of HAN, we first provide an overview of this previous work and discuss its drawbacks that we aim to address.

### A. Problem Definition

In real world scenario, data objects of different types and various interactions between them can be formed into a heterogeneous information network [1].

*Definition 1:* Heterogeneous information network (HIN) is defined as $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ where $\mathcal{V}$ represents a node set of multiple types, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents an edge set of multiple types, and $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$ represents a node attribute matrix. Moreover, $\mathcal{V}$ consists of $m$ types of nodes: $\mathcal{V}_1 = \{v_{11}, \cdots, v_{1n_1}\}$, $\cdots$, $\mathcal{V}_m = \{v_{m1}, \cdots, v_{mn_m}\}$, where $v_{ij}$ represents the $j$-th instance of type $i$.

In this paper, we target on the semi-supervised classification task of nodes in type $\mathcal{V}_1$ without loss of generality. Suppose the node type $\mathcal{V}_1$ contains $n$ nodes of $C$ classes. For each node $v_{1i} \in \mathcal{V}_1$, we have an associated d-dimensional feature vector $\vec{x}_i \in \mathbb{R}^d$ and a class label variable $y_i \in \{1, \cdots, C\}$. We would like to infer the labels $y_i$ for a subset of $v_{1i} \in \mathcal{V}_1$.

### B. Hierarchical Attention Mechanism

The hierarchical attention mechanism proposed by HAN [11] gives a good interpretation of the semantic information at both the node level and the meta-path level. It consists of the following stages.

*1) Meta-path Expansion:*

*Definition 2:* Meta-path $\Phi$ is a path defined as $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_{m-1}} A_m$, abbreviated as $A_1 A_2 \dots A_m$. The relation $R$ between $A_1$ and $A_m$ is a composite relation $R = R_1 \circ R_2 \circ \dots R_m$ where $\circ$ denotes the composition operator on relations.

Meta-path expansion, which finds meta-path based neighbors for each node of the target type, transforms the original heterogeneous information network into several homogeneous sub-graphs containing the target type only.

*2) Node-wise Attention:* Within the meta-path based neighborhood, each node plays a distinct role and exhibits varying degrees of importance in learning the embedding of the target node. Therefore, a self-attention mechanism is adopted here to learn the node-wise attention between the target node and its meta-path based neighbors.

To examine the similarity between nodes, a meta-path $\Phi_i$ specific linear transformation $\mathbf{W}^{\Phi_i,l} \in \mathbb{R}^{F' \times F}$ of neural layer $l$ is used to project the input node features $\vec{h}^{\Phi_i,l} \in \mathbb{R}^F$ to the same embedding space. Then a single feedforward neural layer $att^{\Phi_i,l} : \mathbb{R}^{F'} \times \mathbb{R}^{F'} \to \mathbb{R}$ is shared in computing the attention coefficient as follows:

$$\alpha_{v,v'}^{\Phi_i,l} = \text{softmax}_{v'}(att^{\Phi_i,l}(\mathbf{W}^{\Phi_i,l}\vec{h}_v^{\Phi_i,l}, \mathbf{W}^{\Phi_i,l}\vec{h}_{v'}^{\Phi_i,l})) \tag{1}$$

which indicates the importance of the meta-path based neighbor $v' \in \mathcal{N}^{\Phi_i}(v)$ to the target node $v$ [10]. After that, the meta-path specific embedding of node $v$ in the next neural layer $l+1$ can be computed through the linear aggregation of the products of $v$'s attention to $v'$ and $v'$'s projected features. A nonlinear activation function is applied here.

$$\vec{h}_v^{\Phi_i,l+1} = \sigma\big(\sum_{v' \in \mathcal{N}^{\Phi_i}(v)} \alpha_{v,v'}^{\Phi_i,l}\mathbf{W}^{\Phi_i,l}\vec{h}_{v'}^{\Phi_i,l}\big) \tag{2}$$

*3) Semantic-wise Attention:* Different meta-path based sub-graphs represent distinct semantic relationships. To obtain $v$'s final embedding in the entire heterogeneous information network, we need to ensemble $v$'s semantic information across meta-paths. Each node should value each semantic relationship in a disparate way when learning the final embedding. Hence, the semantic-wise attention should be calculated separately for each node.

To learn the importance of each meta-path specific to each node, we compute the dot-product similarity between the attention vector $\vec{W}_\beta^{\Phi_i} \in \mathbb{R}^{F'}$ for meta-path $\Phi_i$ and node $v$'s final embedding $\vec{H}_v^{\Phi_i} \in \mathbb{R}^{F'}$ in the corresponding meta-path based sub-graph:

$$\beta_v^{\Phi_i} = \text{softmax}_{\Phi_i}\big(\vec{W}_\beta^{\Phi_i} \cdot \vec{H}_v^{\Phi_i}\big) \tag{3}$$

Here, $\vec{\beta}_v^{\Phi_i}$ represents the semantic contribution of meta-path $\Phi_i$ to the final embedding of node $v$ in the entire heterogeneous information network. With the learned semantic-wise attention $\vec{\beta}_v^{\Phi_i}$ as coefficients, linear fusion is carried out by aggregating node $v$'s semantic specific embedding $\vec{H}_v^{\Phi_i}$ so as to generate its final embedding $\vec{Z}_v$.

$$\vec{Z}_v = \sum_{i=1}^{I} \vec{\beta}_v^{\Phi_i} \cdot \vec{H}_v^{\Phi_i} \tag{4}$$

For semi-supervised node classification, we aim to minimize the cross-entropy loss over labeled nodes $\mathcal{V}'$:

$$L = -\sum_{v \in \mathcal{V}'} \vec{Y}_v \ln(\mathbf{Q} \cdot \vec{Z}_v) \tag{5}$$

where $\vec{Y}_v$ is the one-hot vector indicating node $v$'s label, and $\mathbf{Q}$ is the classifier parameter.

### C. Drawbacks

*1) Localized Nature:* As proved in [12], Graph Neural Network is a special form of Laplacian smoothing, which aligns the attributes of a node with its neighbors. Repeatedly applying Laplacian smoothing may mix the features of nodes from different clusters and make them indistinguishable. This drawback is aggravated in the case of HIN. Meta-path expansion can cause the number of edges in the meta-path based homogeneous sub-graph to increase exponentially. When nodes are densely connected, the mixing happens dramatically fast as we stack more neural layers. Therefore, the number of node-wise attention layers used in the hierarchical attention mechanism is limited. However, since a shallow neural network cannot sufficiently propagate the label information from the training set to the entire graph, the hierarchical attention mechanism suffers from the localized nature of the node-wise attention.

| Notation | Description |
|---|---|
| $I$ | Number of meta-paths |
| $\Phi_i$ | Meta-path $i$ |
| $L$ | Number of node-wise attention layers |
| $\mathbf{W}^{\Phi_i,l}$ | Projection matrix in layer $l$ for meta-path $i$ |
| $\alpha_{v,v'}^{\Phi_i,l}$ | Attention to neighbor $v'$ in layer $l$ for meta-path $i$ |
| $\vec{h}_v^{\Phi_i,l}$ | Node $v$'s embedding in layer $l$ for meta-path $i$ |
| $\vec{W}_\beta^{\Phi_i}$ | Attention vector of meta-path $i$ |
| $\beta_v^{\Phi_i}$ | Node $v$'s attention to meta-path $i$ |
| $\mathbf{H}^{\Phi_i}$ | Embedding matrix for meta-path $i$ |
| $\mathcal{U}^{\Phi_i}$ | Virtual non-local node set for meta-path $i$ |
| $\hat{\mathbf{X}}^{\Phi_i}$ | Non-local feature matrix for meta-path $i$ |
| $\mathcal{N}^{\Phi_i}(v)$ | Meta-path $i$ based neighbors of node $v$ |
| $w_{v,v'}^{\Phi_i}$ | Edge weight between $v$ and $v'$ for meta-path $i$ |
| $\mathcal{B}^{\Phi_i,l}$ | Sampled node set in layer $l$ for meta-path $i$ |
| $\mathbf{Z}$ | Final embedding matrix |

TABLE I: Notation table

Nonetheless, [14] points out that capturing long-range dependencies is of primary importance in deep neural networks. Apparently, each node-wise attention layer in HAN only allows nodes to aggregate information from their direct neighbors. The generally shallow structure of the node-wise attention mechanism limits the model's ability to attend to long-range dependencies in the graph.

*2) Computational Inefficiency:* The edge density of each meta-path based sub-graph is associated with that of the original heterogeneous information network and the property of the corresponding meta-path. If the original HIN is densely connected or the selected meta-path is relatively long, meta-path expansion can give rise to a dramatic increase in edge density.

As shown in Table II, the number of authors is 14K in the DBLP dataset and most of the authors only connect to a few number of papers. The original graph is relatively sparsely connected. However, with the expansion based on meta-path APCPA which represents a more composite relation, the number of edges in the resulting sub-graph increases to around 19M in total. Since the calculation of the node-wise attention examines all neighbor pairs in the meta-path based sub-graph, a huge increase in edge density due to meta-path expansion in datasets like DBLP can make the calculation computationally infeasible.

## IV. Large-scale Non-local Learning in HIN

To address the drawbacks discussed in Section III-C, we propose to extend the hierarchical attention mechanism to large-scale non-local learning in HIN. More specifically, we handle the localized nature of the mechanism via a non-local attention structure in Section IV-A and IV-B. Then the node-wise weighted sampling schema and jumping knowledge aggregation are used to reduce the computational inefficiency in Section IV-C and IV-D.

### A. Non-local Attention Structure

To capture non-local dependencies in HIN, we propose to inject those dependencies into the shallow structure of the hierarchical attention mechanism. On the one hand, this will not cause oversmoothness due to the stacking of neural layers. On the other hand, our framework can still attend to information beyond local neighborhood even with a shallow structure. In this way, both local and non-local features can be captured at the same time.

In comparison to [14], we compute long-range dependencies in the graph by calculating structural similarities between any two nodes in the graph, regardless of their positional distance. Besides, in order to maintain the semantic integrity of each meta-path, we only carry out non-local operation within each meta-path based sub-graph.

The generic non-local information calculation is defined as

$$\hat{\vec{x}}_v^{\Phi_i} = \frac{1}{C(\vec{x}_v)} \sum_{v' \in \mathcal{V}^{\Phi_i}} f(v,v') g(\vec{x}_{v'}) \tag{6}$$

where $\vec{x}_v$ and $\hat{\vec{x}}_v^{\Phi_i}$ are the target node $v$'s input features and generated meta-path $\Phi_i$ based non-local features, respectively. $v'$ enumerates all nodes in the same sub-graph. $f$ is a non-local measure used to compute a scalar similarity between $v$ and $v'$, $g$ gives a task specific transformation of $\vec{x}_v$, and $C(\vec{x}_v^{\Phi_i})$ is the normalization factor.

Then we create a virtual node $u_v^{\Phi_i}$ to contain non-local features $\hat{\vec{x}}_v^{\Phi_i}$ and set it as a neighbor to the target node $v$ in the meta-path $\Phi_i$ based sub-graph. This virtual node $u_v^{\Phi_i}$ connects to the target node $v$ only. The node-wise attention learns the importance of $u_v^{\Phi_i}$ and $v$'s meta-path $\Phi_i$ based neighborhood together.

$$\vec{h}_v^{\Phi_i,l+1} = \sigma\big( \sum_{v' \in \mathcal{N}^{\Phi_i}(v) \cup \{u_v^{\Phi_i}\}} \alpha_{v,v'}^{\Phi_i,l} \mathbf{W}^{\Phi_i,l} \vec{h}_{v'}^{\Phi_i,l} \big) \tag{7}$$

Here, $v'$ enumerates over the union of $v$'s meta-path $\Phi_i$ based neighborhood $\mathcal{N}^{\Phi_i}(v)$ and the virtual neighbor $u_v^{\Phi_i}$. When $v' = u_v^{\Phi_i}$, we set $\vec{h}_{v'}^{\Phi_i,l} = \hat{\vec{x}}_v^{\Phi_i,l}$. This process allows $v$ to attend to both local and non-local information at the same time and determine the usefulness of each piece of information via training.

At the stage of node-wise attention, we do not wish to mix the non-local information with local information. Hence, the virtual node $u_v^{\Phi_i}$'s features $\hat{\vec{x}}_v^{\Phi_i,l}$ in Layer $l$ are transformed to the next embedding space only via the projection matrix $\mathbf{W}^{\Phi_i,l}$:

$$\hat{\vec{x}}_v^{\Phi_i,l+1} = \sigma\big( \mathbf{W}^{\Phi_i,l} \cdot \hat{\vec{x}}_v^{\Phi_i,l} \big) \tag{8}$$

To have a deeper insight into this non-local attention structure, we give an illustration in Figure 3. By isolating the virtual node $u_v^{\Phi_i}$'s contribution to the target node $v$'s embedding $\vec{h}_{v'}^{\Phi_i}$, we have

$$\begin{aligned} \alpha_{v,u_v^{\Phi_i}}^{\Phi_i} \mathbf{W}^{\Phi_i} \hat{\vec{x}}_v^{\Phi_i} &= \alpha_{v,u_v^{\Phi_i}}^{\Phi_i} \mathbf{W}^{\Phi_i} \sum_{v' \in \mathcal{V}^{\Phi_i}} f(v,v') g(\vec{x}_{v'}) \\ &= \sum_{v' \in \mathcal{V}^{\Phi_i}} \alpha_{v,u_v^{\Phi_i}}^{\Phi_i} f(v,v') \mathbf{W}^{\Phi_i} g(\vec{x}_{v'}) \end{aligned} \tag{9}$$

Therefore, by attending to the non-local features contained in the virtual neighbor $u_v^{\Phi_i}$, the target node $v$ essentially attends

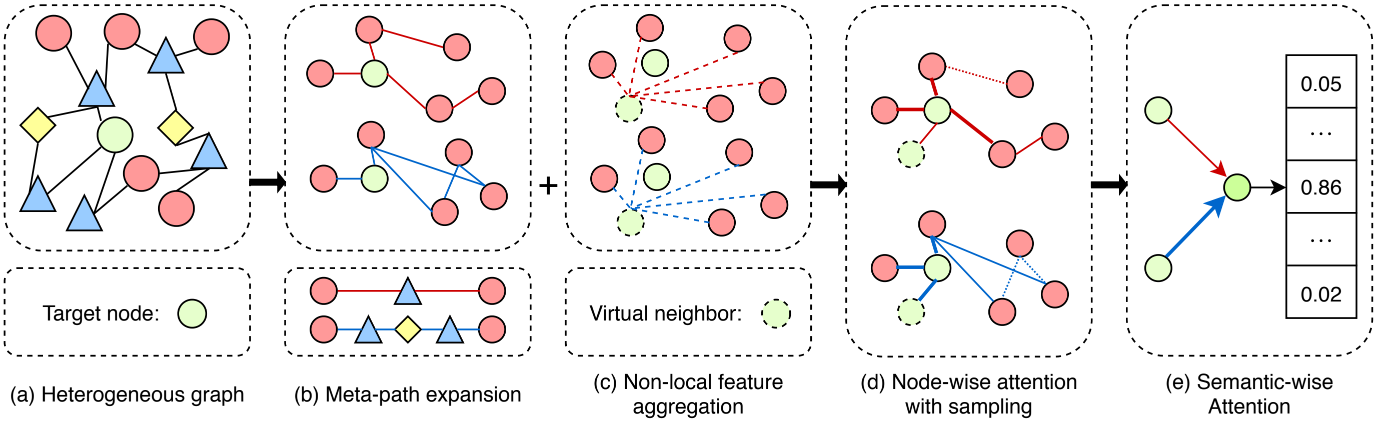

Target node:
Virtual neighbor:
+
0.05
...
0.86
...
0.02
(a) Heterogeneous graph
(b) Meta-path expansion
(c) Non-local feature aggregation
(d) Node-wise attention with sampling
(e) Semantic-wise Attention

Fig. 2: The overall framework. (a) The original heterogeneous information networks. (b) Homogeneous sub-graphs after meta-path expansion using two different meta-paths. (c) Non-local feature aggregation for the target node to create virtual neighbors. (d) Node-wise attention with sampling. A thicker edge represents a higher attention. (e) Semantic-wise attention to generate final embedding.

to all nodes within the meta-path $\Phi_i$ based sub-graph. This serves the purpose of capturing long-range dependencies. The non-local attention between the target node $v$ and any node $v' \in \mathcal{V}^{\Phi_i}$ is the product of the attention $\alpha^{\Phi_i}_{v,u_v^{\Phi_i}}$ between $v$ and $u_v^{\Phi_i}$ and the precomputed similarity $f$ between $v$ and $v'$.

$$\alpha^{\Phi_i}_{v,v'} = \alpha^{\Phi_i}_{v,u_v^{\Phi_i}} \cdot f(v,v') \ , \ \forall v' \in \mathcal{V}^{\Phi_i} \tag{10}$$

In this way, the non-local attention between $v$ and any node $v'$ is an approximation of the true attention that $v$ would pay to $v'$ if there was a real edge between them. However, it is computationally infeasible to allow each node to really pay attention to all other nodes during training. This kind of approximation, therefore, saves a lot of computation power as the similarity $f(v,v')$ is precomputed only once before training. Each node can then determine the balance between local and non-local information via the node-wise attention mechanism.

Since the non-local attention between $v$ and any node $v'$ simulates the case that there was a real edge between them, the receptive field of $v$ essentially exceeds the scope of $v$'s direct neighborhood even with a limited number of attention layers. Meanwhile, this kind of simulation also enriches the semantic meaning and densifies the original graph. This is especially important to those relatively sparsely connected sub-graphs.

### *B. Non-local Measures*

Based on Equation 10, $\alpha^{\Phi_i}_{v,u_v^{\Phi_i}}$ is shared by all nodes in the computation of the non-local attention $\alpha^{\Phi_i}_{v,v'}$. Hence, the similarity generated by the non-local measure $f(v,v')$ should be approximately proportional to the true attention that $v$ would pay to $v'$ if there was a real edge between them. In other words, $f(v,v')$ must capture the latent semantic relationships beyond local neighborhood. We propose three different non-local measures $f(v,v')$ that can fulfill this requirement and fit our non-local attention structure. Here, $f(v,v')$ is defined within each meta-path based sub-graph.

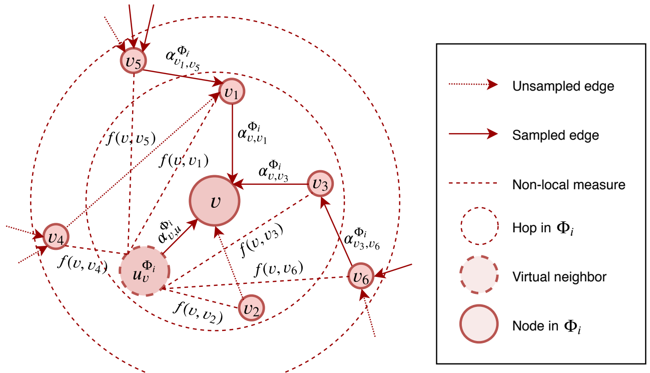


Fig. 3: An illustration of non-local structure.

**Second-order Proximity.** The second-order proximity makes an assumption that nodes with similar distributions over the context tend to have similar semantic roles in the graph [21]. Let $\vec{S}_v = (w_{v,1}, \ldots, w_{v,|\mathcal{V}|})$ define the first-order proximity between $v$ and all other nodes, where $w_{v,v'}$ is the weight of the edge $(v,v')$. Then the second-order proximity between $v$ and $v'$ is determined by the similarity between $\vec{S}_v$ and $\vec{S}_{v'}$. Softmax is used for normalization.

$$f(v,v') = \text{softmax}_{v'}(\vec{S}_v^{\top} \cdot \vec{S}_{v'}) \tag{11}$$

**Personalized PageRank.** Personalized PageRank [29] measures the importance of other nodes $v'$ relative to the root node $v$ based on the underlying assumption that more important nodes are likely to receive more edges from other nodes. Using the symmetrically normalized adjacency matrix $\hat{\tilde{\mathbf{A}}} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{A} \tilde{\mathbf{D}}^{-\frac{1}{2}}$, where $\mathbf{A}$ is the adjacency matrix, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ and $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$, we have:

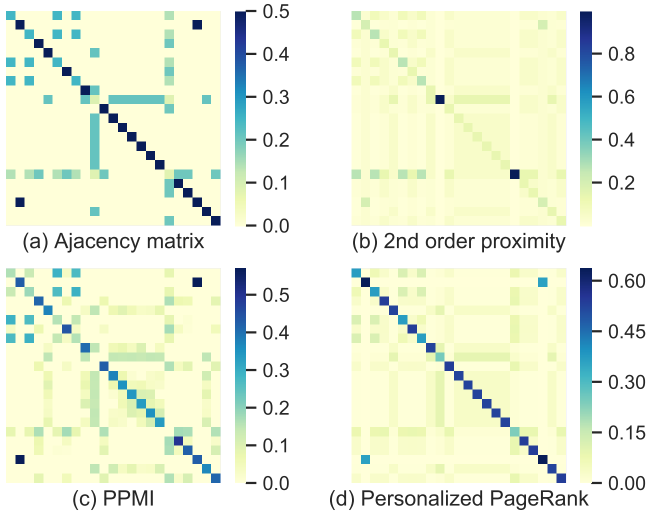
$$f(v,v') = \left[\alpha(\mathbf{I} - (1-\alpha)\hat{\tilde{\mathbf{A}}})^{-1}\right]_{v,v'} \tag{12}$$

where $[\cdot]_{v,v'}$ represents the $(v,v')$ entry of the matrix $\cdot$ and $\alpha \in (0,1]$ is the restart probability.

**Positive Pointwise Mutual Information.** Positive pointwise mutual information encodes the estimated probability that node $v$ occurs in context $v'$ [22], [30]. If there is a semantic relation between $v$ and $v'$, then $f(v, v')$ is expected to be greater than if $v$ and $v'$ are independent. We first use random walk to generate a co-occurence frequency matrix $\mathbf{F}$, then calculate the non-local measure as:

$$f(v, v') = \max\{\log(\frac{\mathbf{F}_{v,v'}}{\mathbf{F}_{v,*}\mathbf{F}_{*,v'}}), 0\} \tag{13}$$

where $\mathbf{F}_{v,*}$ and $\mathbf{F}_{*,v'}$ are the sum of row $v$ and column $v'$ in $\mathbf{F}$, respectively.



(a) Ajacency matrix

(b) 2nd order proximity

(c) PPMI

(d) Personalized PageRank

Fig. 4: Heatmap of normalized matrices.

Figure 4 visualizes the matrices constructed by running the above three different non-local measures on a sub-graph of the DBLP network. Compared with the adjacency matrix of the same sub-graph, there are two obvious differences: the effect of the hub nodes is reduced; more latent long-range dependencies are captured via the non-local measures. This suggests that the proposed non-local measures meet the requirement raised at the beginning of this section.

In this section, we propose three possibilities, but the choice of non-local measures used in NLAH framework is not limited to them. Any measures that can capture latent long-range dependencies in the graph can be applied here. To make a more complete evaluation, we compare the performance of the proposed three non-local measures and investigate the reasons why certain measures perform better in certain datasets in Section V-C1.

### C. Weighted Sampling Schema

As discussed in Section III-C2, meta-path expansion typically gives rise to a huge increase in edge density. Since the computation cost of the node-wise attention is proportional to the number of edges in the meta-path based sub-graph, a high edge density can cause an unaffordable computation cost. Hence, we propose a novel sampling method which is specially designed for large-scale heterogeneous information networks. Existing works [8], [20], [31] target on homogeneous networks and sample nodes based on the graph structure. However, we wish to leverage the rich semantic information in heterogeneous information networks, and therefore, sample nodes based on the information density of each edge and each meta-path based sub-graph.

**Algorithm 1:** NLAH Algorithm

```
Input  : Heterogeneous information network
         $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$,
         Meta-path set $\Phi$, Target node set $\mathcal{B}$,
         Number of layers $L$
Output: Trained model, Final embeddings $\mathbf{Z}$
for $\Phi_i \in \Phi$ do
    Generate meta-path based sub-graph
      $G^{\Phi_i} = (\mathcal{V}^{\Phi_i}, \mathcal{E}^{\Phi_i})$;
    Generate virtual neighbors $\mathcal{U}^{\Phi_i}$ with non-local
      features $\hat{\mathbf{X}}^{\Phi_i}$ using Equation 6;
    $\mathcal{B}^{\Phi_i,L} \leftarrow \mathcal{B}$;
    for $l \leftarrow \mathcal{L} - 1$ to $0$ do
        $\mathcal{B}^{\Phi_i,l} \leftarrow \mathcal{B}^{\Phi_i,l+1}$;
        for $v \in \mathcal{B}^{\Phi_i,l}$ do
            Weighted sampling $\mathcal{N}^{\Phi_i,l}(v)$ from $\mathcal{N}^{\Phi_i}(v)$
              using Equation 15;
            $\mathcal{B}^{\Phi_i,l} \leftarrow \mathcal{B}^{\Phi_i,l} \cup \mathcal{N}^{\Phi_i,l}(v)$;
        end
    end
    $\vec{h}_v^{\Phi_i,0} \leftarrow \vec{x}_v$ where $v \in \mathcal{B}^{\Phi_i,0}$ and $\vec{x}_v \in \mathbf{X}$;
    $\vec{\hat{x}}_v^{\Phi_i,0} \leftarrow \vec{\hat{x}}_v^{\Phi_i}$ where $v \in \mathcal{B}^{\Phi_i,0}$ and $\vec{\hat{x}}_v^{\Phi_i} \in \hat{\mathbf{X}}^{\Phi_i}$;
    for $l \leftarrow 0$ to $\mathcal{L} - 1$ do
        for $v \in \mathcal{B}^{\Phi_i,l+1}$ do
            Learn the attention $\alpha_{v,v'}^{\Phi_i,l}$ using Equation 1
              where $v' \in \mathcal{N}^{\Phi_i,l}(v) \cup \{u_v^{\Phi_i}\}$;
            Learn embedding $\vec{h}_v^{\Phi_i,l+1}$ using Equation 7;
            Get $\vec{\hat{x}}_v^{\Phi_i,l+1}$ using Equation 8;
        end
    end
    Get $\mathbf{H}^{\Phi_i}$ from $\{\mathbf{H}^{\Phi_i,1}, \mathbf{H}^{\Phi_i,2} \ldots \mathbf{H}^{\Phi_i,L}\}$ using
      Equation 16;
end
Learn the weight $\beta^{\Phi_i}$ of meta-path $\Phi_i$ using Equation
  3;
Fuse the semantic embeddings to generate final
  embeddings $\mathbf{Z}$ using Equation 4;
Calculate the task-specific loss using Equation 5 and
  back propagate to update parameters;
```

The number of neighbors sampled for each node in a meta-path based sub-graph is formulated as

$$\max\left(\ln\left(d^{\Phi_i}\right), n\right) \tag{14}$$

Here, $d^{\Phi_i}$ is the average node degree for the meta-path $\Phi_i$ based sub-graph and $n$ is a hyperparameter. The general idea is that a higher average node degree suggests richer semantic information, hence more neighbors are needed to capture the information. But to avoid an extreme degree, we normalize all degrees by natural log. The $n$ is the lower bound in case that the graph is too sparse.

If multiple instances of the same meta-path exist between two nodes, there is a high chance that these two nodes share a high semantic similarity. Then we assign the number of meta-path instances to the weight of the edge between these two nodes. The probability of sampling the neighbor $v'$ for $v$ should be proportional to the edge weight $w_{v,v'}^{\Phi_i}$:

$$p(v'|v, \Phi_i) = \frac{w_{v,v'}^{\Phi_i}}{\sum_{v'' \in \mathcal{N}^{\Phi_i}(v)} w_{v,v''}^{\Phi_i}} \tag{15}$$

Since there is a higher probability of sampling semantically similar nodes, the weighted sampling schema helps to reduce the sampling variance as shown in Figure 8. In this way, our framework can be more effectively and efficiently applied to many big data applications.

### *D. Jumping Knowledge Aggregation*

The weighted sampling schema discussed in the last section samples a fixed number of neighbors for each node. However, this limits the ability of our model in capturing the diverse local structure of sub-graphs. As a result, we adopt the jumping-connection idea [18] to combine a node's intermediate representations from each node-wise attention layer.

$$\vec{H}_v^{\Phi_i} = \max\left(\vec{h}_v^{\Phi_i,1}, \vec{h}_v^{\Phi_i,2}, \ldots \vec{h}_v^{\Phi_i,L}\right) \tag{16}$$

We conduct element-wise max pooling over node $v$'s intermediate embeddings $\vec{h}_v^{\Phi_i,l}$ from layer $l$. This approach ensembles information across layers and better captures the diverse local structure.

### *E. Analysis of Proposed Model*

The overall process is summarized in Algorithm 1. We give an illustration of our framework in Figure 2 and analyze it as follows:

- The proposed framework of non-local attention in heterogeneous information networks (NLAH) captures long-range dependencies in HIN through the non-local attention structure, which aids the local semantic features learned by the hierarchical attention mechanism. The learned attention is highly interpretable and benefits graph analysis.
- The overall attention structure can be efficiently implemented through parallelization. Weighted sampling schema makes the computation more efficient in real-world large-scale datasets. The time complexity of one node-wise attention layer is $O(VFF' + UF)$ and that of the semantic-wise attention layer is $O(VF'I)$, where $V$ and $U$ are the number of nodes and sampled nodes, respectively.

## V. Experiments

The goal of our experiments is to examine the effectiveness and efficiency of the proposed NLAH framework.

| Dataset | #Node | Meta-path | #Meta instance | Feature | Label |
|---|---|---|---|---|---|
| DBLP | 14K | APA | 40K | 300 | 4 |
| | | APCPA | 19M | | |
| ACM | 3K | PAP | 13K | 1870 | 3 |
| | | PSP | 1M | | |
| SLAP | 20K | GTG | 303K | 2695 | 15 |
| | | GDG | 7K | | |
| | | GPG | 416K | | |
| | | GG | 172K | | |
| | | GDCDG | 18K | | |

TABLE II: Summary of dataset statistics.

### *A. Experiment Settings*

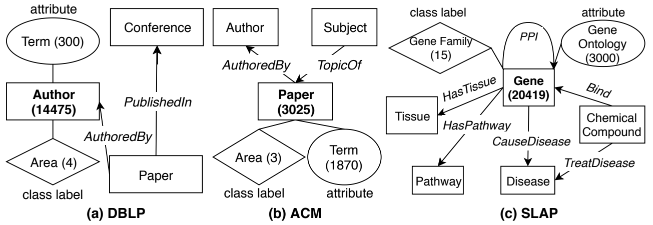


Fig. 5: Dataset schema.

*1) Datasets:* We apply our framework to 3 different types of heterogeneous information networks [2], and the statistics are summarized in Table II.

1) **DBLP:** We use the bibliographic information network extracted from DBLP dataset, which forms a bipartite network. It includes 3 types of nodes: Author, Paper and Conference, connected via 2 types of edges: AuthoredBy and PublishedIn. We choose Author as the target instance and the network schema is shown in Figure 5. We extract a bag-of-words representation (300 words) of all the paper abstracts published by an author as his or her input features and assign each author to a research area.
2) **ACM:** We extract papers from different conferences in the ACM dataset and construct a star schema network accordingly. It consists of 3 types of nodes: Paper, Author and Subject, and 2 types of edges: AuthoredBy and TopicOf. We choose Paper as the target instance and demonstrate the network schema in Figure 5. Each paper's bag-of-words representation (1870 words) and research area are used as instance attributes and labels, respectively.
3) **SLAP:** We use the multiple-hub network in a bioinformatic dataset SLAP [32]. It contains node types such as chemical compound, gene, disease, and pathway etc. We choose gene as the target instance and use 15 gene families as instance labels. The network schema is shown in Figure 5. 3000 gene ontology terms (GO terms) are extracted for each gene as instance features.

*2) Baselines:* We compare our framework NLAH with some state-of-the-art baselines, including GNN-based methods and random walk based methods. For methods [10], [22], [23] targeting homogeneous networks, we test them on each meta-path based sub-graph and report the best performance.

1) **GAT** [10]: A GNN-based method which utilizes the attention mechanism in homogeneous networks.
2) **APPNP** [23]: A GNN-based method which constructs a personalized propagation of neural predictions in homogeneous networks.
3) **DGCN** [22]: A GNN-based method which uses two convolutional networks to jointly consider local and global consistency in homogeneous networks.
4) **metapath2vec** [26]: A random walk based method which performs meta-path based random walks in HIN.
5) **ESim** [33]: A random walk based method which uses user guidance to capture semantic information from multiple meta-paths in HIN.
6) **HAN** [11]: A GNN-based method which considers hierarchical local attention in HIN.
7) **NLAH**$_{\text{2ndprox}}$: The proposed framework which uses both the hierarchical attention mechanism and the non-local attention structure. Here we use the second-order proximity as the non-local measure.
8) **NLAH**$_{\text{ppr}}$: The proposed framework which uses the personalized PageRank as the non-local measure.
9) **NLAH**$_{\text{ppmi}}$: The proposed framework which uses the positive point-wise mutual information as the non-local measure.

*3) Implementation Details:* For all GNN-based models (GAT, APPNP, DGCN, HAN, NLAH), we experiment with 1 to 5 layers and report the best performance. During training, we practice the train-validation-test split with a ratio of 3:1:1 for each dataset and use early stopping with a patience of 50. Each model is optimized with Adam with a learning rate of 0.005, a regularization parameter of 0.0005 and a dropout rate of 0.4. More specifically, for the proposed NLAH framework, we set $g(\vec{x}_{v'})$ in Equation 6 as the identity function and $n = 1$ in Equation 14. For GAT and HAN, we use 8 attention heads. For APPNP, we set restart probability $\alpha = 0.1$. For random walk based methods (metapath2vec, ESim), we set window size to 5, walk length to 50, walks per node to 50, the number of negative samples to 5. To ensure fairness, an embedding dimension of 64 is adopted for all the above algorithms.

### B. Evaluation

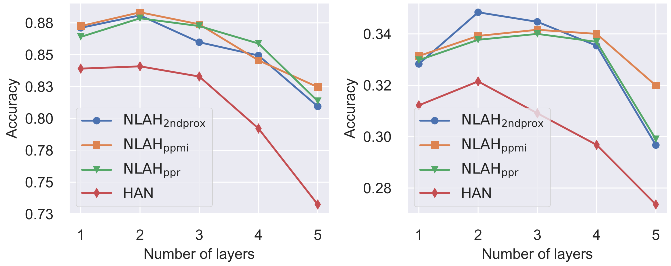
To thoroughly examine the effectiveness of the proposed method, we evaluate NLAH against other state-of-the-art baselines on two representative tasks: node classification and embedding visualization.

*1) Node Classification:* We first want to see how the proposed NLAH framework performs in comparison to the existing state-of-the-art methods. Here we employ linear SVM classifiers for the node classification task. Each model is tested under the same experiment setting 10 times and the average results are summarized in Table III. NLAH achieves the best performance in all three datasets. More specifically, NLAH using the second order proximity as the non-local measure obtains the highest results in DBLP and SLAP datasets, while the one using the positive pointwise mutual information outperforms the rest in ACM dataset. We can also observe that the performance of methods targeting homogeneous networks (GAT, APPNP, DGCN) is largely affected by the edge density of meta-path based sub-graphs. Hence, their node classification results vary across datasets. In particular, NLAH is ranked the top using all metrics since the incorporation of non-local features include additional knowledge which cannot be captured by stacking more local attention layers (GAT, HAN).

| Dataset | DBLP | | ACM | | SLAP | |
|---|---|---|---|---|---|---|
| Metrics | Acc | F1 | Acc | F1 | Acc | F1 |
| GAT | 93.73 | 92.51 | 85.55 | 85.45 | 17.47 | 21.18 |
| APPNP | 73.92 | 61.36 | 85.60 | 85.50 | 30.91 | 23.99 |
| DGCN | 75.40 | 68.55 | 84.56 | 84.39 | 18.62 | 20.56 |
| metapath2vec | 87.33 | 84.64 | 74.96 | 75.74 | 33.61 | 6.29 |
| ESim | 92.37 | 91.11 | 73.48 | 74.73 | 31.78 | 6.10 |
| HAN | 91.02 | 89.70 | 84.09 | 84.87 | 32.15 | 26.57 |
| NLAH$_{\text{2ndprox}}$ | **96.92** | **96.48** | 88.09 | 88.09 | **34.84** | 28.80 |
| NLAH$_{\text{ppr}}$ | 96.56 | 95.95 | 87.86 | 87.90 | 34.00 | 28.51 |
| NLAH$_{\text{ppmi}}$ | 96.43 | 95.91 | **88.33** | **88.33** | 34.16 | **29.10** |

TABLE III: Node classification results (%).

*2) Layer Experiment:* Since the proposed NLAH framework aims to resolve the problems caused by the localized nature of HAN, we conduct a layer experiment on NLAH and HAN by gradually increasing the number of node-wise attention layers used for the node classification task in ACM and SLAP datasets. The setting in Section V-B1 is adopted and the results are drawn in Figure 6. Both models' performances peak at two or three layers. When more layers are added, the classification accuracy is impaired significantly by oversmoothness. In addition, we also notice that NLAH with one layer, regardless of the non-local measures used, always beats the best performance achieved by HAN. This phenomenon coincides with our analysis in Section III-C1. Even though HAN can attend to information beyond the direct neighborhood by stacking more attention layers, the drawback of oversmoothness associated with an increasing number of layers offsets the performance boost brought by extra information. In contrast, when only one layer is used in NLAH, the problem of oversmoothness is negligible. Nonetheless, NLAH can still inject non-local information into its shallow structure to achieve a higher accuracy.

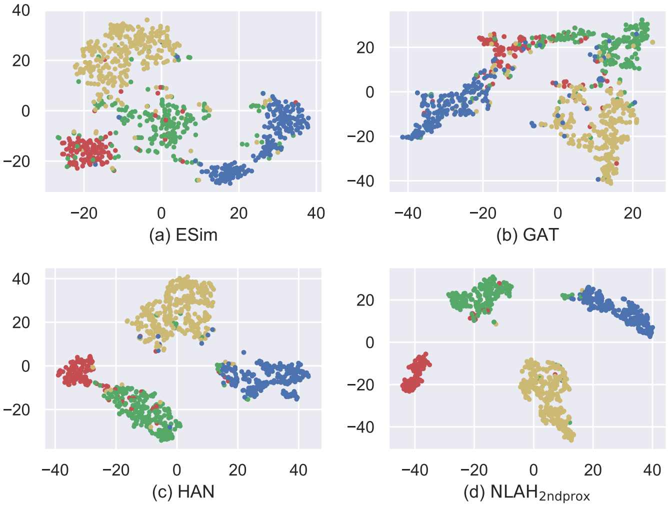


**(a) ACM** **(b) SLAP**

Fig. 6: Layer experiment.

*3) Embedding Visualization:* To give a more intuitive comparison, we visualize the learned node embeddings on a 2-

dimensional space using t-SNE [34]. 4 different models are used to embed author nodes in DBLP test set and the results are given in Figure 7. Since GAT is designed for homogeneous networks, it cannot well separate nodes with different labels. Due to the nature of random walk, ESim is more likely to mix the embeddings of nearby nodes even if they are from different clusters. HAN performs better as its separate treatment of each meta-path preserves the semantic integrity. NLAH clearly achieves the best performance by forming distinct groups for nodes with different labels. It not only ensembles semantic meanings carried by each meta-path, but also attend to non-local features simultaneously. As a result, node embedding learned by NLAH exhibits high intra-class similarity and inter-class difference.



(a) ESim

(b) GAT

(c) HAN

(d) NLAH$_{2ndprox}$

Fig. 7: Node embedding visualization for DBLP test set using t-SNE. Different colors indicate different labels.

### *C. Analysis*

*1) Comparison of the Proposed Non-local Measures:* We aim to compare the three proposed non-local measures in terms of their performance in different datasets. By cross-referencing the results in Table III and the heatmaps in Figure 4, we can see that the personalized PageRank is the least capable of capturing latent long-range dependencies, thus its performance is relatively lower than the other two non-local measures. The performance of positive pointwise mutual information is often affected by the hyperparameters of the random walk used in the generation process. It is relatively harder to fully explore a large graph with only a few steps of random walk. As a result, it performs better in a smaller dataset like ACM. On the other hand, the performance of the second order proximity is not affected by the graph size. Hence, it gives the highest results in both DBLP and SLAP datasets.

*2) Non-local Attention Structure as a Form of Graph Densification:* As discussed in Section IV-A, since the non-local attention structure allows each node to attend to all other nodes, this process can be viewed as a form of graph densification. To prove this point, we gradually remove certain percentage of edges in ACM dataset and run NLAH against other baselines on the resulting datasets. We stop at removing 60% of edges since removing any more edges can give rise to isolated nodes and render the node classification task meaningless. The setting in Section V-B1 is adopted and the results are summarized in Table IV.

| Edges Removed | 0% | 20% | 40% | 60% |
|---|---|---|---|---|
| GAT | 85.55 | 84.99 | 84.38 | 83.48 |
| APPNP | 85.60 | 85.08 | 84.32 | 83.85 |
| DGCN | 84.56 | 82.92 | 83.48 | 81.88 |
| metapath2vec | 74.96 | 74.27 | 74.10 | 72.75 |
| ESim | 73.48 | 72.12 | 67.88 | 65.62 |
| HAN | 84.09 | 82.86 | 82.41 | 81.95 |
| NLAH$_{2ndprox}$ | 88.09 | 87.23 | 86.34 | 86.80 |
| NLAH$_{ppr}$ | 87.86 | 87.62 | **87.39** | 86.85 |
| NLAH$_{ppmi}$ | **88.33** | **87.91** | 87.09 | **86.87** |

TABLE IV: Node classification accuracy (%) after removing certain percent of edges in ACM dataset.

NLAH outperforms the rest on all the resulting datasets. In particular, NLAH using the positive pointwise mutual information achieves the highest accuracy in most cases. More importantly, the performance of NLAH using all three types of non-local measures is more resistant to the removal of edges. Since GNN-based methods suffer from localization, when more edges are removed, each node can only attend to less neighbors. Consequently, the model extracts less features from the local neighborhood and gives deteriorated performance. In contrast, the non-local attention structure in NLAH captures latent semantic relationships even if two nodes are not directly linked. The incorporation of non-local information alleviates the negative effects caused by the drop in the amount of local information. This suggests that the improvement in performance by NLAH is more significant in sparse graphs.

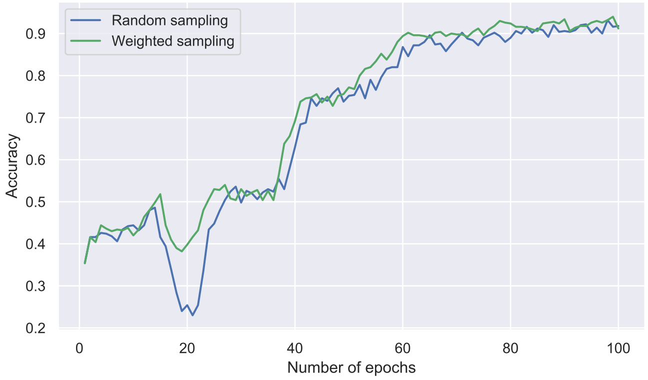


Fig. 8: Performance curves of the model trained using two different sampling schemas.

*3) Variance Reduction due to Weighted Sampling Schema:* A further analysis is conducted to examine if weighted sampling of meta-path based neighbors can truly reduce the variance during the training process. We run the NLAH model on the DBLP dataset using weighted and random sampling. The training curves based on the node classification results of the validation set are drawn in Figure 8. The model trained using weighted sampling presents a more smooth performance curve as compared with the one using random sampling. The reduction in training variance is significant, especially at around Epoch 20, where the drop in accuracy is largely regularized by weighted sampling.

## VI. Conclusion

In this paper, we propose a novel framework of non-local attention in heterogeneous information networks (NLAH) for semi-supervised HIN representation learning. The framework utilizes the non-local attention structure to complement the hierarchical attention mechanism. In this way, it allows each node in the graph to simultaneously leverage both local semantic information and long-range dependencies. A properly designed weighted sampling schema is deployed to improve computational efficiency. Empirical studies on three different types of heterogeneous information networks demonstrate the effectiveness of the NLAH framework. Although we focus on studying its impacts on node-related tasks, the proposed technique is very general and can be applied to many other tasks such as link-related tasks (e.g. link prediction) and application-related tasks (e.g. social recommendation). Exploration of those other tasks will be an interesting direction for future research. Meanwhile, we also seek to investigate how other types of non-local measures can fit our framework and how other types of heterogeneous information network embedding methods can be generalized under our framework in the future.

## References

[1] Y. Sun and J. Han, "Mining heterogeneous information networks: principles and methodologies," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 3, no. 2, pp. 1–159, 2012.

[2] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2017.

[3] S. Bhagat, G. Cormode, and S. Muthukrishnan, "Node classification in social networks," in *Social network data analytics*. Springer, 2011, pp. 115–148.

[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*. ACM, 2016, pp. 855–864.

[5] Z. Liu, V. W. Zheng, Z. Zhao, F. Zhu, K. C.-C. Chang, M. Wu, and J. Ying, "Semantic proximity search on heterogeneous graph by proximity embedding," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 7, no. 3, p. 11, 2013.

[7] C. Yang, C. Zhang, X. Chen, J. Ye, and J. Han, "Did you enjoy the ride? understanding passenger experience via heterogeneous network embedding," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 2018, pp. 1392–1403.

[8] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, 2017, pp. 1024–1034.

[9] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[11] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *Proceedings of the 2019 World Wide Web Conference (WWW)*, 2019.

[12] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] G. Taubin, "A signal processing approach to fair surface design," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. ACM, 1995, pp. 351–358.

[14] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[15] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *International conference on machine learning*, 2016, pp. 2014–2023.

[16] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in neural information processing systems*, 2016, pp. 3844–3852.

[17] S. Abu-El-Haija, B. Perozzi, R. Al-Rfou, and A. A. Alemi, "Watch your step: Learning node embeddings via graph attention," in *Advances in Neural Information Processing Systems*, 2018, pp. 9180–9190.

[18] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," *arXiv preprint arXiv:1806.03536*, 2018.

[19] H. Gao and H. Huang, "Self-paced network embedding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1406–1415.

[20] J. Chen, T. Ma, and C. Xiao, "Fastgcn: fast learning with graph convolutional networks via importance sampling," *arXiv preprint arXiv:1801.10247*, 2018.

[21] 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*. International World Wide Web Conferences Steering Committee, 2015, pp. 1067–1077.

[22] C. Zhuang and Q. Ma, "Dual graph convolutional networks for graph-based semi-supervised classification," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018, pp. 499–508.

[23] J. Klicpera, A. Bojchevski, and S. Günnemann, "Predict then propagate: Graph neural networks meet personalized pagerank," 2018.

[24] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[25] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu, "Relevance search in heterogeneous networks," in *Proceedings of the 15th international conference on extending database technology*. ACM, 2012, pp. 180–191.

[26] 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*. ACM, 2017, pp. 135–144.

[27] T.-y. Fu, W.-C. 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*. ACM, 2017, pp. 1797–1806.

[28] Y. Zhang, Y. Xiong, X. Kong, S. Li, J. Mi, and Y. Zhu, "Deep collective classification in heterogeneous information networks," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018, pp. 399–408.

[29] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[30] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[31] W. Huang, T. Zhang, Y. Rong, and J. Huang, "Adaptive sampling towards fast graph representation learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 4558–4567.

[32] B. Chen, Y. Ding, and D. J. Wild, "Assessing drug target association using semantic linked data," *PLoS computational biology*, vol. 8, no. 7, p. e1002574, 2012.

[33] J. Shang, M. Qu, J. Liu, L. M. Kaplan, J. Han, and J. Peng, "Meta-path guided embedding for similarity search in large-scale heterogeneous information networks," *arXiv preprint arXiv:1610.09769*, 2016.

[34] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.