Computing and Informatics, Vol. 44, 2025, 635–660, doi: 10.31577/cai_2025_3_635

# ADAPTIVE SAMPLING-BASED HETEROGENEOUS GRAPH ENHANCEMENT

Qin Zhao, Guojun Yang, Yaru Miao, Jie Lian, Hongda Qi*

*Shanghai Egineering Research Center of Intelligent Education and Big Data*
*Shanghai Normal University, Shanghai 200234, China*
*e-mail:* hongda_qi@shnu.edu.cn

Zuliang Kou

*Shanghai Newtouch Software Co., Ltd.*
*Shanghai 200127, China*
*e-mail:* zuliang.kou@newtouch.com

**Abstract.** In the current research on heterogeneous academic network community detection, there is a widespread challenge of high demand for node representation of node attributes in learning graphs. Particularly, existing methods often perform poorly when dealing with nodes missing attributes. Furthermore, most methods rely on meta-paths, but the optimal length of meta-paths is difficult to determine and the quality of predefined meta-paths directly affects the results. To address this issue, this paper proposes an Adaptive Sampling-based Heterogeneous Graph Enhancement Model (ASGNN). The model aims to solve the problem of inaccurate node representations leading to imprecise community partitions in academic networks. ASGNN first effectively captures the network's topological structure through random walk techniques, and then utilizes an adaptive sampling algorithm to select the most influential adjacent node set, rather than relying on traditional meta-path techniques. The model further employs an attention mechanism to aggregate information from nodes of different types, thereby enhancing attribute completion and topological structure in heterogeneous academic networks. This approach not only fills in missing information but also significantly enhances the semantic and structural integrity of the network. Experimental results demonstrate that the proposed model exhibits outstanding performance on two real datasets compared to baseline models.

* Corresponding author

**Keywords:** Heterogeneous graphs, missing attribute, adaptive sampling, attribute completion, topological augmentation

**Mathematics Subject Classification 2010:** 68-T30

## 1 INTRODUCTION

In the real world, the complex relationships among various entities can be modeled as graphs, such as social networks, biological molecule networks, network communication graphs, academic networks, etc. In recent years, with the development of technology and the advent of the "era of academic big data," academic networks have emerged [1]. Academic networks encompass not only citation relationships among papers but also "writing" relationships between authors and papers, "publishing" relationships between papers and journals, "co-authorship" relationships among authors, and so on [2]. The diverse types of nodes and relationships make heterogeneous academic networks rich in information. Scholars' exploration of academic networks originated in the field of library and information science [3], primarily for the evaluation of scientific literature. With the advancement of academic networks in scientific research, scholars have gradually begun to investigate academic networks. Community detection in academic networks not only reveals the affiliation information of academic papers and discovers research hotspots but also facilitates recommending academic papers on the same topic to students, as well as multi-topic recommendations [4, 5, 6]. Existing research on community detection in academic networks mainly falls into two categories: homogeneous academic network community detection [7] and heterogeneous academic network community detection [8]. Homogeneous academic network community detection only considers paper nodes in the network and the "citation relationships" between them. Heterogeneous academic networks, on the other hand, consider multiple types of nodes and relationships. Compared to homogeneous academic networks, heterogeneous academic networks cover richer information, such as author and venue information. This poses challenges to existing technologies while assisting the primary nodes in obtaining deeper structural and semantic insights. Heterogeneous graphs containing rich structural information increase the difficulty of representation learning. This paper primarily addresses this challenge by studying representation learning in heterogeneous academic networks. The concept of heterogeneous networks was proposed by Sun et al. [9] in 2009. They argued that heterogeneous graphs are a special type of network that includes multiple types of objects, and meaningful results cannot be obtained without considering the types of network nodes. Real-world networks are heterogeneous, and studying academic networks in a single dimension cannot comprehensively explore their underlying structures.

When studying heterogeneous networks, it is often challenging to model higher-order relationships between nodes. Most existing research focuses on homogeneous

graphs, although there are some metrics that can serve heterogeneous graphs, most of which rely on predefined meta-paths [10, 11, 12]. Meta-paths generally require experienced domain experts to define, and the quality of predefined meta-paths directly affects the final results. The variety of meta-paths in heterogeneous networks poses a challenge in effectively selecting and utilizing relevant meta-paths. Additionally, different meta-paths may express the same or similar semantics, introducing the issue of information redundancy [13]. Determining the optimal length of meta-paths is difficult; overly long paths may introduce noise, while overly short paths may fail to express complete semantics. The selected meta-paths cannot cover all semantics in the network, resulting in information loss. The use of meta-paths tends to overly emphasize node types while overlooking important relationships between nodes of the same type. Heterogeneous graph Attention Network (HAN) [14] utilizes meta-paths to obtain higher-order information of nodes. It performs attention calculations on subgraphs generated from nodes to aggregate neighbor information and further calculates attention between different types to aggregate information. Some methods do not use meta-paths; Luo et al. [15] proposed a heterogeneous graph embedding method based on contextual paths, avoiding the need for directly using meta-paths by adaptively generating contextual paths. Although this method effectively avoids the challenge of needing predefined meta-paths, the generated contextual paths inherently suffer from the problem of not covering all semantics in the network and overly emphasizing node types while ignoring relationships between nodes. Heterogeneous Graph Neural Network (HetGNN) [16] selects heterogeneous nodes through sampling to avoid directly using meta-paths. However, in the sampling process, the initial number of sampled nodes for different types ignored the contributions of different types of nodes to the primary type nodes.

Despite the numerous methods based on Graph Neural Networks (GNNs) being used to model graph data [17, 18, 19], the performance of traditional GNN methods remains unsatisfactory for pattern-rich heterogeneous graphs. To address this issue, recent methods focus on heterogeneous networks have been proposed, such as HAN, Metapath Aggregated Graph Neural Network (MAGNN) [20], etc. These methods can be understood as ways of analyzing node attributes guided by the graph structure. However, learning node representations from the graph in this manner places high demands on node attributes, and some nodes often lack attributes, possibly due to the high cost of acquisition and concerns about privacy, among other reasons. In heterogeneous graphs, it is typically impossible to obtain attribute information for all types of nodes, posing a challenge to the model's performance. Additionally, due to limitations of datasets and challenges in data integration, relationships between nodes may also be missing. For the sake of illustration, we categorize all types of nodes in the graph into two types: primary type nodes, which are the nodes of interest for study, and auxiliary type nodes, which assist in the analysis. In theory, nodes of any type can be considered primary type nodes. Taking the DBLP dataset as an example, suppose we aim to perform community detection in an academic network, considering "paper" nodes as primary type nodes. However, due to the difficulty in obtaining their attributes, as well as the attributes of auxiliary type nodes that

have a significant impact on "paper" nodes, the effectiveness of academic network analysis is significantly compromised. Heterogeneous academic networks suffer from issues such as missing attribute information and improper neighbor selection. How to select appropriate domain nodes and enhance the network? Existing graph enhancement methods mainly focus on structural enhancement of the graph, ignoring attribute enhancement of nodes. Even when considering attribute enhancement, it typically involves simply replacing missing attributes with the mean value, which fails to fully consider the uniqueness of node attributes, resulting in model smoothing.

Most of the aforementioned methods require predefined meta-paths, which limit the application of heterogeneous graphs greatly, as predefined meta-paths necessitate significant prior knowledge. Moreover, the quality of defined meta-paths directly impacts experimental results. To address this issue, this paper proposes a heterogeneous graph enhancement model based on adaptive sampling to enhance the expressiveness of GNNs on heterogeneous networks. When primary type nodes miss attribute information, the model adaptively selects neighboring nodes with strong relevance for information aggregation, thus improving the attribute information of primary type nodes, as illustrated in Figure 1.

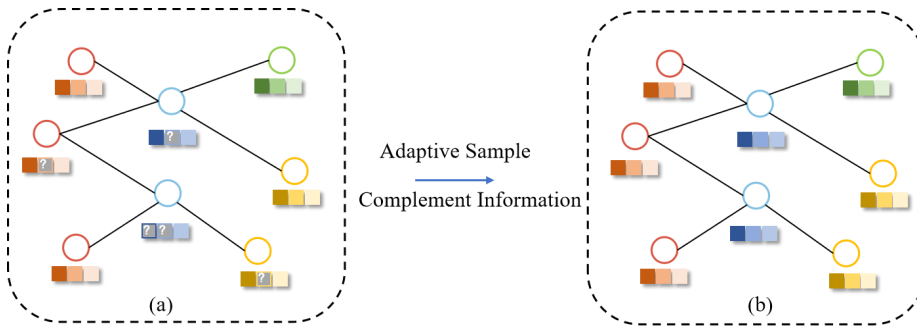



Figure 1. Attribute completion

Therefore, we first execute a random walk strategy on the subgraph of the target node to capture the graph's topology. Then, adaptive sampling is performed on the node to obtain a highly correlated heterogeneous neighbor node set through sampling probability. Next, feature enhancement is conducted on primary type nodes. Specifically, considering dynamically extracting node features to capture the correlation among same-type neighboring nodes and obtain deeper information, we first employ Bidirectional Long Short-Term Memory (BiLSTM) [21] to extract features from the same-type node set. Subsequently, based on attention mechanisms, feature aggregation is performed on different types of node sets, flexibly addressing the relationships between different types of nodes in heterogeneous networks to more accurately capture the semantic relationships of the network. Finally, the concept of virtual edges is introduced. By computing the average correlation between

different types of nodes and the target node and comparing it with the correlation between node pairs, connections between nodes are increased to enhance the overall topological structure performance. This enables better utilization of heterogeneous academic networks for downstream tasks. In summary, the main contributions of our work are as follows:

- Based on the heterogeneous academic network, we propose an adaptive sampling method for sampling primary and auxiliary type node pairs. The sampling weights of different auxiliary type nodes are determined based on the relevance of the sampled node pairs. Subsequently, nodes of different types are sampled according to their respective sampling weights.
- When enhancing the attributes of primary type nodes, we combine BiLSTM with attention mechanisms to capture the correlation among same-type neighboring nodes while also flexibly addressing the relationships between nodes of different types.
- By comparing the relevance score between all nodes of a certain type and the target node with the average relevance score of the induced n-order subgraph of that type of node on the target node, we determine whether to add virtual edges for structural enhancement.
- Compared to the baseline model, our proposed model demonstrates superior performance across multiple real datasets.

The structure of this paper is as follows: Section 2 reviews related work on heterogeneous graph embedding learning, Section 3 defines the problem, Section 4 provides a detailed introduction to our algorithm, Section 5 discusses the experimental results. Finally, Section 6 concludes the paper.

## 2 RELATED WORK

Heterogeneous graph representation learning aims to generate meaningful vector representations for each node while preserving the heterogeneous structure and semantics for downstream tasks such as node/graph classification, node clustering, and link prediction. Kaibiao et al. [22] proposed a method that defines neighborhoods through different levels of sampling methods and utilizes neighborhood graphs to represent complex structural interactions between nodes. A hierarchical attention mechanism was employed to learn the importance of different objects. Zhu et al. [23] introduced a method for automatic attribute completion, enhancing the performance of heterogeneous GNN models through differential attribute completion algorithms and dual-layer optimization techniques. Zhang et al. [24] introduced a method that enriches learned node embeddings with both structural information and attribute semantics. They improved the final performance through finely-tuned multi-relation aggregation modules and multi-layer convolutional modules. Zhao et al. [25] proposed a method that achieves more refined node embeddings and enhances community detection in heterogeneous information networks through enhanced neighbor

selection and structured perception information aggregation. He et al. [26] developed an unsupervised heterogeneous graph contrastive learning approach (HGCA) to handling attribute missingness in heterogeneous information networks (HINs). Based on metapath-based random walk techniques, it learns joint embeddings of nodes and attributes to achieve fine-grained attribute completion. Jin et al. [27] proposed a generic framework based on attribute completion, which utilizes existing HIN-embedding methods to compute attention coefficients between nodes. Attribute completion for nodes without attributes is achieved through weighted aggregation. Fu et al. [28] proposed the Structural Enhanced Graph Convolutional Network (SEGCN), which introduces a structural enhancement technique, integrating network structure, attribute information, and higher-order neighborhood relations to achieve a more comprehensive node representation. Li et al. [29] proposed a method for attribute completion through feature completion. They designed a heterogeneous residual graph attention network to learn the graph's topology and then utilized attention mechanisms to complete missing features.

Some methods [30, 31] model heterogeneous graphs using predefined meta-paths. Zhang et al. [32] proposed seven categories of "intra-network social meta-paths" and four categories of "inter-network social meta-paths." These "social meta-paths" cover various connection information in the network and help solve multi-network link prediction problems. Wang et al. [14] first introduced HAN, which is based on hierarchical attention, including node-level and semantic-level attention. Node-level attention aims to learn the importance of nodes and their neighbors based on meta-paths, while semantic-level attention learns the importance of different meta-paths. The proposed model generates node embeddings by hierarchically aggregating features from meta-path-based neighbors. MAGNN [20] defines multiple meta-paths in heterogeneous graphs to capture compound relationships and guide neighbor selection. In the selected meta-paths, MAGNN utilizes all node information along the meta-paths, unlike HAN, which ignores intermediate nodes in meta-paths. Drawing inspiration from Generative Adversarial Networks (GANs), Hu et al. [33] introduced HeGAN, a novel framework for HIN-embedding that trains a discriminator and a generator in a minimax game. Wang et al. [34] proposed a approach that differs from previous approaches focused on node-level or relation-level heterogeneity modeling by jointly integrating the rich semantics retained on nodes and relations for modeling. Zhong et al. [35] proposed a new embedding model that aims to reduce the dependence of heterogeneous network models on manually defined meta-paths by selecting task-relevant meta-paths through automatic mining. Hu et al. [36] proposed a method that considers structural information, semantic information, metapath-based node features, and weights based on metapaths to learn effective node embeddings. They employed a relation-aware heterogeneous graph neural network (GNN) to generate compact embeddings for nodes. Zhang et al. [37] introduced a novel definition of metapaths that integrates edge types (i.e., relationships between nodes), and utilizes different subgraphs to separately train node embeddings, aggregating nodes from different subgraphs using attention mechanisms. Lou et al. [38] proposed a method that, through graph data augmentation and adaptive denoising

mechanisms, is capable of uncovering hidden relationships between heterogeneous nodes and enhancing information propagation between them, while being robust to variations in graph structure and noise. Liu et al. [39] proposed the Metapath-based Multi-level Graph Attention Network (MMAN), which jointly learns node embeddings on heterogeneous graphs and performs node classification and clustering on two types of substructures.

However, some of the aforementioned methods either rely on meta-paths for node information aggregation or do not consider adaptive sampling. This can lead to the model not fully utilizing the semantic and structural information of heterogeneous networks, resulting in decreased model performance.

## 3 PROBLEM STATEMENT

**Definition 1.** Heterogeneous academic network graph (HG) definition. Given HG, $G = (V, E, T, \phi, \varphi)$, where $V$ and $E$ are sets of nodes and edges, respectively. Each node $v$ and edge $e$ are associated with their type mapping functions $\phi : V \to T_V$ and $\varphi : E \to T_E$, where $T_V$ and $T_E$ represent sets of node and edge types. The types satisfy $|T_V| + |T_E| > 2$, and $T = T_V \cup T_E$. If $|T_V| + |T_E| = 2$, then the graph has only one type of node and one type of edge, which will degenerate into a homogeneous graph. The primary type nodes studied in this paper are paper nodes, theoretically nodes of any type can be considered primary type nodes. Node types are classified as $\{\mathcal{P}, \mathcal{A}, \mathcal{T}, \mathcal{V}\}$, representing Paper, Author, Term, and Venue, respectively.

The edge types are classified as $\{r_1, r_2, r_3\}$, representing the following relationships: $r_1$ denotes the "relate to" relationship between a $\mathcal{P}$ and the $\mathcal{T}$ it addresses; $r_2$ indicates the "write" relationship between an $\mathcal{A}$ and a $\mathcal{P}$; and $r_3$ represents the "publish in" relationship between a $\mathcal{P}$ and the $\mathcal{V}$ where it is published. The feature matrix of the graph is denoted as $H$, where $h_i^{T_i}$ represents the initial feature representation of node $v_i$ of type $T_i$. The adjacency matrix is denoted as $A$, where $a_{ij} = 1$ indicates that nodes $i$ and $j$ are connected. Figure 2 illustrates the information of the heterogeneous academic network.

**Problem Definition.** Given HG, $G = (V, E, T, \phi, \varphi)$, the objective of this paper is to project nodes into a latent low-dimensional representation space while obtaining an enhanced graph $G'$. Formally, our goal is to study a mapping function $f$ such that $(G', X_V) = f(G, H_V)$, where $X_V$ represents the final embedding of the nodes.

## 4 ASGNN FRAMEWORK

In this section, we will formally introduce ASGNN, designed to address the challenges of missing attribute information for certain nodes and missing edge information between nodes in heterogeneous graphs. ASGNN consists of four components:

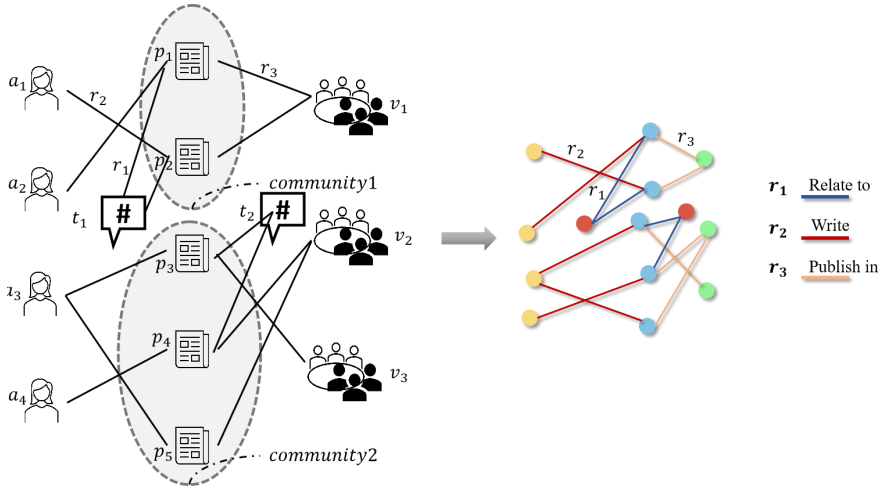



Figure 2. Heterogeneous academic network

1. Adaptive sampling of heterogeneous neighbors,
2. Feature extraction,
3. Feature aggregation,
4. Topology enhancement.

Figure 3 illustrates the framework of ASGNN.

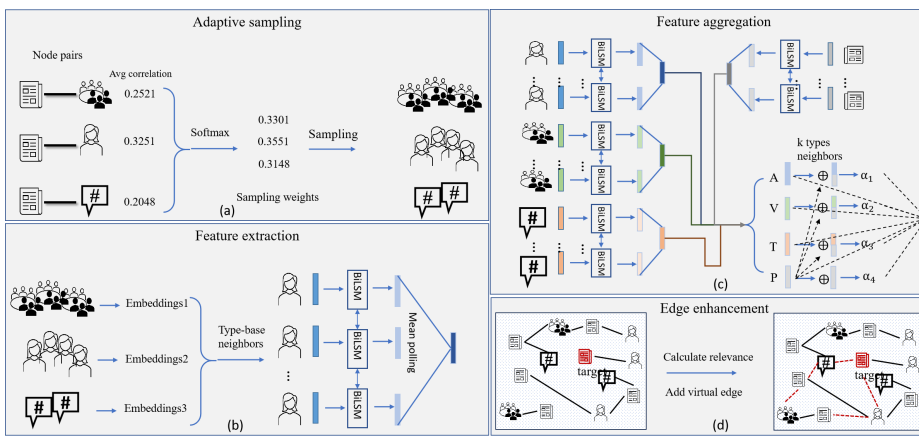



Figure 3. ASGNN

### 4.1 Heterogeneous Graph Adaptive Sampling

Many researchers have proposed various methods to aggregate information for nodes in heterogeneous graphs. However, when selecting neighboring nodes, most methods rely solely on nodes with strong attribute similarity, ignoring structural relationships between nodes. According to researchers' findings [40], the structural correlation between nodes is highly significant for modeling node similarity. In this paper, we choose to adopt adaptive sampling to sample neighborhood nodes strongly correlated with each target node.

Firstly, random walks are conducted in the induced subgraph of each node $v$, sampling neighboring nodes. To avoid randomness and compute the influence of different types of neighboring nodes on the current node equally, we choose to sample the same number of neighboring nodes for each different type. The correlation between the target node and neighboring nodes of different types is calculated based on the initial feature representation of nodes. Firstly, the correlation between the target node and each neighbor (denoted as $\mathcal{P}$ and $\mathcal{A}$ for illustration) is calculated using cosine similarity. In this formula, $Coll_{\mathcal{P}_i,\mathcal{A}_j}$ represents the correlation between node $i$ of type $\mathcal{P}$ and node $j$ of type $\mathcal{A}$:

$$Coll_{\mathcal{P}_i,\mathcal{A}_j} = \cos(\theta) = \frac{h_i^{\mathcal{P}} \cdot h_j^{\mathcal{A}}}{\|h_i^{\mathcal{P}}\| \cdot \|h_j^{\mathcal{A}}\|}, \tag{1}$$

where $h_i^{\mathcal{P}}$ and $h_j^{\mathcal{A}}$ represent the embeddings of nodes $i$ of type $\mathcal{P}$ and $j$ of type $\mathcal{A}$, respectively.

To measure the influence of neighboring nodes of different types on the target node, we calculate the correlation between the sampled neighboring nodes of different types and the target node, and take the average among all types of neighboring nodes. This ensures that the contribution of nodes of each type is considered during information aggregation. Specifically, $Coll_{\mathcal{P},\mathcal{A}}^{AVG}$ represents the average correlation between nodes of type $\mathcal{P}$ and nodes of type $\mathcal{A}$. The calculation of the average correlation between different types of neighbors and the target node is as follows:

$$Coll_{\mathcal{P},\mathcal{A}}^{AVG} = \frac{\text{sum}\left(Coll_{\mathcal{P}_i,\mathcal{A}_j}\right)}{\text{count}(\mathcal{A}_j)}. \tag{2}$$

To determine the relative importance of neighboring nodes of different types on the correlation with the target node, we employ a *softmax* normalization strategy. By applying *softmax* normalization to the average correlation between nodes of heterogeneous types and the target node, we obtain the influence of each type of neighboring node on the target node. $W_t$ is defined as the weight matrix, and its calculation is as follows:

$$W_t = Softmax\left(Coll_{\mathcal{P},t}^{AVG}\right) = \frac{e^{Coll_{\mathcal{P},t}^{AVG}}}{\sum_{c=1}^{c} e^{Coll_{\mathcal{P},c}^{AVG}}}, \quad t \in \{\mathcal{A}, \mathcal{T}, \mathcal{V}\}. \tag{3}$$

The weight matrix $W_t$ obtained through calculations serves as the adaptive sampling weight for each type of heterogeneous neighboring node, enabling targeted selection of neighboring nodes. This weight allocation mechanism ensures that node types with a greater impact on the target node are more frequently sampled, improving the accuracy and representativeness of node representations obtained during the information aggregation process. As illustrated in Figure 3 a), taking the analysis of paper nodes as an example, suppose we calculate the average correlations between authors, venue, term, and paper nodes to be 0.3251, 0.2521, and 0.2048, respectively. In this scenario, we sample neighboring nodes based on the calculated weights mentioned above. For instance, given a certain number of sampling iterations, we sample according to the weights of author nodes (0.3551), venue nodes (0.3301), and term nodes (0.3148), ensuring a more accurate reflection of the contribution of each type of neighboring node to the target node during the aggregation of neighbor information.

### 4.2 Feature Extraction Based on BiLSTM

After sampling the neighboring nodes, the next crucial step is to aggregate features of these sampled nodes. We first need to aggregate information among neighboring nodes of the same type, as nodes of the same type typically possess similar features and semantics. Leveraging these similarities helps to better capture the collective contribution of neighboring nodes of the same type to the target node.

For the target node $v \in V$, we focus on its sampled neighboring nodes of type $t$, denoted as $v' \in N_t(v)$. By aggregating these $t$-type neighboring nodes through a neural network, we can better capture their correlations and interactions. The process of information aggregation is as follows:

$$f_1^t(v) = AG^t_{v' \in N_t(v)}\{h^t_{v'}\}, \tag{4}$$

where $f_1^t$ is the vector resulting from aggregating information of $t$-type neighbors, $h^t_{v'}$ represents the initial feature representation of the $t$-type node $v'$, and $AG^t$ denotes the aggregation function for nodes of $t$-type.

The paper employs BiLSTM for information aggregation, allowing bidirectional reading of node information and dynamically extracting node features to obtain deeper insights, as shown in Figure 3 b). Additionally, using BiLSTM for aggregating node information helps alleviate the vanishing gradient problem. This information aggregation aims to capture the correlations among neighboring nodes of the same type, making the representation of the target node more expressive. Therefore, Equation (4) can be rewritten as:

$$f_1^t(v) = \frac{\sum_{v' \in N_t(v)} \left[\overrightarrow{LSTM}\{h^t_{v'}\} \bigoplus \overleftarrow{LSTM}\{h^t_{v'}\}\right]}{|N_t(v)|}, \tag{5}$$

training different BiLSTM networks separately to handle nodes of different types in

the heterogeneous academic network allows for a more comprehensive capture of the unique features and complex relationships of each node type. This separate training process enables each node type to receive personalized treatment, enhancing the model's expressive power for each node type and thereby improving the performance of graph enhancement tasks.

### 4.3 Feature Aggregation Based on Attention Mechanism

In the preceding sections, the target node's features are aggregated with the features of each type of neighboring node, and $O_m$ represents the set of types for the target node and its neighboring nodes. For each type of neighboring node, we perform separate aggregation, and we then further integrate these different aggregated vectors to obtain a more comprehensive representation of the target node.

Considering that different types of neighboring nodes contribute differently to the target node $v$, we introduce an attention mechanism to jointly learn the contributions of different node types to the target node. This mechanism adaptively determines the weight of each node type in the information aggregation, allowing for a more flexible capture of the importance of different types of neighboring nodes in the heterogeneous academic network, as shown in Figure 3 c). The output embedding of node $v$ is represented as $x_v$:

$$x_v = \alpha^{v,v} h_v^{\mathcal{P}} + \sum_{t \in O_m} \alpha^{v,t} f_1^t(v), \tag{6}$$

where $\alpha^{(v,*)}$ represents the importance of nodes of $*$-type to the current node, $h_v^{\mathcal{P}}$ is the initial embedding of node $v$, and $f_1^t(v)$ represents the heterogeneous neighbor embedding based on type.

To more effectively aggregate the neighbor information of the target node while preserving the node's own information, this paper designs different weights for information aggregation. We define the set of embeddings as $F(v) = \{f_0^t(v), f_1^t(v), \ldots, f_{|O_m|-1}^t(v), f_0^t(v) = h_v^t, t \in O_m \subset T_V\}$, where $f_i^t(v)$ represents the representation of the $t$-type neighbor nodes sampled for node $v$. We can rewrite Equation (6) as:

$$x_v = \sum_{f_i \in F(v)} \alpha^{v,i} f_i, \tag{7}$$

the acquisition of attention coefficients $\alpha^{v,i}$ is as follows:

$$\alpha^{v,i} = \frac{\exp\{LeakyReLU(u^T[f_i] \oplus h_v^t)\}}{\sum_{f_i \in F(v)} \exp\{LeakyReLU(u^T[f_i] \oplus h_v^t)\}}, \tag{8}$$

where $u^T$ is a vector used to map $[f_i] \oplus h_v^t$ to a scalar value. After obtaining the aggregation weights for different types of neighbor nodes with the target node, we obtain the final representation $x_v$ for the target node.

### 4.4 Enhancement Based on Sampled Virtual Edges

In heterogeneous academic networks, enhancing node features alone is not sufficient. This is because the edge relationships between nodes are likely to be missing, and these relationships have a significant impact on the model's expressiveness. Therefore, to better capture the relationships between nodes, we need to enhance the network topology. For example, we can consider enhancing the network topology in two aspects. Theme Association: By adding virtual edges between paper nodes and term nodes, we can better represent the semantic relationships between paper topics and term. This helps improve the model's understanding of paper content, thus more accurately capturing the distribution of topics in the academic network. Collaboration Relationships: Adding virtual edges between author nodes represents collaboration relationships between authors. This enables the model to better understand the social network of authors, thereby better exploring the impact of collaboration relationships on academic research.

We determine whether to add virtual edges by comparing the correlation between a certain type of node and the target node with the average correlation score of that type of node in the induced $n$-order subgraph of the target node. Specifically, we follow these steps:

1. Correlation Calculation: For the target node $v$ and all nodes $u_i^t \in \{u_i \mid \phi(u_i) = T_t\}$ of a certain type, we compute the correlation score between them, reflecting the strength of their connection.
2. Average Correlation Score Calculation: Compute the average of the correlation scores obtained for all nodes of the same type in the $n^{\text{th}}$ order subgraph induced by the target node.
3. Decision for Adding Virtual Edge: Compare the correlation score between the target node and a node of a certain type with the average correlation score. If the correlation score is greater than or equal to the average correlation score, add a virtual edge to strengthen the connection between them; otherwise, if the correlation score is less than the average correlation score, do not add a virtual edge.

This topological enhancement strategy is adjusted based on specific tasks and network characteristics, thereby it is more flexibly adapting to the analytical requirements of academic networks in different scenarios. First, compute the correlation score between the target node $v$ and all nodes $u_i^t$ of a certain type:

$$Coll_{v,u_i^t} = \cos(\theta) = \frac{h_v \cdot h_{u_i^t}}{||h_v|| \cdot ||h_{u_i^t}||}, \tag{9}$$

where $h_v$ represents the embedding of node $v$, $h_{u_i^t}$ represents the embedding of node $u_i^t$, $Coll_{v,u_i^t}$ represents the correlation between node $v$ and the $t$-type node $u_i^t$. The decision to add an edge between node $v$ and node $u_i^t$ depends on the value of $judge$.

An edge is added between node $v$ and node $u_i^t$ if and only if $judge = 1$. The calculation rule for $judge$ is as follows:

$$judge = \begin{cases} Coll_{v,u_i^t} \geq Coll_{T_v,T_t}^{AVG}, & 1, \\ Coll_{v,u_i^t} < Coll_{T_v,T_t}^{AVG}, & 0, \end{cases} \tag{10}$$

the nodes pairs with $judge = 1$, determined through computation, were used to enhance the topology of the heterogeneous academic network. On the left side of Figure 3 d), the original academic network is depicted. By calculating the correlations between all nodes of the main type (Paper) and nodes of other types, and comparing the correlations between nodes and their respective types, we identified the edges that needed enhancement. This process resulted in the enhanced topology graph shown on the right side of Figure 3 d).

### 4.5 Algorithmic Description

ASGNN aims to adaptively sample neighbor nodes based on the correlation between main nodes and their neighbors, thereby obtaining more relevant neighbor information even in the absence of node attributes, achieving attribute enhancement. Then, based on the relative importance of the correlations between nodes and the correlations between node types, it decides whether to add new edges between nodes of different types to achieve topological enhancement, thus overall enhancing the performance of academic networks. The algorithmic process of this paper is shown in Algorithm 1.

## 5 EXPERIMENTS

In this section, two evaluation datasets are introduced, and detailed information about the competing algorithms is provided. Then, the focus shifts to downstream node clustering tasks, where we evaluate the performance of the proposed model against other state-of-the-art methods. Finally, ablation studies are discussed to provide further insights.

### 5.1 Datasets

In the experiments, we selected two representative heterogeneous academic network datasets, namely DBLP and ACM. The two datasets contain four and three types of nodes, respectively, along with three types of edges. Their statistics are presented in Table 1.

1. DBLP [14]: The dataset used in this experiment is a subset extracted from the DBLP database, consisting of a set of diverse academic resources. It includes 14 328 research papers authored by 4 057 authors from 20 different academic

**Algorithm 1:** Algorithm for Enhancing Heterogeneous Academic Network Graph

**Input:** Original heterogeneous academic network graph $G$
**Output:** Enhanced heterogeneous academic network graph $G'$ and enhanced representation of primary type nodes $x_v \in \mathcal{P}$

Determine primary type nodes $\mathcal{P}$
**for** $v$ *in* $\mathcal{P}$ **do**
  Construct subgraph $g_v$ for each $v$
  Calculate average relatedness $Coll_{\mathcal{P},t}^{AVG}$ in Equation (2)
  Calculate sampling probability $W_t$ in Equation (3)
  Resample neighbors; node-based neighbors aggregated via BiLSTM, type-based neighbors aggregated via ATTENTION to obtain enhanced representation $x_v$ for primary type nodes in Equations (5), (7) and (8)
**end**
**for** $v$ *in* $\mathcal{P}$ **do**
  Calculate relatedness $Coll_{v,u_i^t}$ between heterogeneous neighbor nodes and the target node in Equation (9)
  Calculate *judge* to determine whether to add a virtual edge in Equation (10)
**end**
**return** $G'$, $x_v \in \mathcal{P}$

venues. The dataset also contains 8 789 terms, where each term represents a fundamental concept discussed in the papers.

2. ACM [14]: This dataset is a subset extracted from the ACM database, consisting of a collection of diverse academic resources. It includes 4 019 research papers authored by 7 167 authors, covering 60 different research subjects. To provide comprehensive representations of papers and authors, the dataset includes various attributes. Each paper's attributes are represented using a bag-of-words approach, effectively capturing terms that define its content essence. Similarly, for authors, their attributes are also represented as bags-of-words, including valuable information extracted from their affiliations, paper titles, and terms from their publications.

### 5.2 Baselines

To validate the effectiveness of the proposed method in this paper, we selected several commonly used baseline algorithms for comparison with our proposed model:

1. DeepWalk [41]: DeepWalk is a network embedding method based on random walks, which combines random walks with the word2vec approach for mining graph data.

| **Dataset** | **Nodes** | **Edges** |
|---|---|---|
| DBLP | #author($\mathcal{A}$): 4 057 | #$\mathcal{A}$-$\mathcal{P}$: 19 645 |
| | #paper($\mathcal{P}$): 14 328 | #$\mathcal{P}$-$\mathcal{T}$: 85 810 |
| | #term($\mathcal{T}$): 7 723 | #$\mathcal{P}$-$\mathcal{V}$: 14 328 |
| | #venue($\mathcal{V}$): 20 | |
| ACM | #paper($\mathcal{P}$): 4 019 | #$\mathcal{P}$-$\mathcal{P}$: 9 516 |
| | #author($\mathcal{A}$): 7 167 | #$\mathcal{P}$-$\mathcal{A}$: 13 407 |
| | #subject($\mathcal{S}$): 60 | #$\mathcal{P}$-$\mathcal{S}$: 4 019 |

Table 1. The statistics of the public datasets

2. GCN [42]: GCN is a neural network architecture used for learning node embeddings in graphs. It effectively captures and aggregates information through convolutional layers by utilizing the neighborhood of the graph. GCN models node relationships and graph structures efficiently, demonstrating significant performance in tasks such as node classification and link prediction.

3. GAT [43]: GAT is a neural network architecture designed specifically for graph-based learning tasks. It employs attention mechanisms to adaptively weigh the importance of neighboring nodes during the information aggregation process. GAT's ability to capture fine-grained dependency relationships in graphs, along with its self-attention mechanism, makes it a powerful tool for various applications, including node classification and graph classification.

4. Metapath2vec [44]: Metapath2vec is a pioneering algorithm for learning embeddings in heterogeneous information networks. It introduces the concept of meta-paths, which are paths composed of multiple node types, to capture both structural and semantic information. By leveraging meta-paths, Metapath2vec can generate rich context-aware embeddings for various types of nodes in heterogeneous networks.

5. HAN [14]: HAN is a cutting-edge model designed for learning on heterogeneous graphs. It combines node-level and metapath-level attention mechanisms to effectively capture both local and global information in heterogeneous graphs. Due to its ability to handle the complexity and diversity of heterogeneous graph data, HAN demonstrates excellent performance in various applications including node classification.

### 5.3 Evaluation Metrics

We employ *ARI* [2] (Adjusted Rand Index) and *NMI* [45] (Normalized Mutual Information) as evaluation metrics. *ARI* is an external index used to assess clustering results. It measures the similarity between two data distributions and is typically used to compare clustering results with the ground truth labels. Its value ranges

from −1 to 1. $ARI$ can be expressed as:

$$\begin{aligned} RI &= \frac{a+b}{\frac{n}{2}}, \\ ARI &= \frac{RI - E[RI]}{\max(E[RI_1], E[RI_2]) - E[RI]}, \end{aligned} \tag{11}$$

where $RI$ is the Rand Index, $E[RI]$ is the expected value of the Rand Index, and $E[RI_1]$ and $E[RI_2]$ are the expected Rand Indexes based on the true labels and clustering results, respectively. $ARI$ provides a more robust and interpretable way of evaluating clustering quality by considering the possibility of random matches.

$NMI$ is a commonly used metric for evaluating clustering performance, measuring the similarity between two clustering results, with a value range from 0 to 1. $NMI$ can be expressed as:

$$\begin{aligned} H(X) &= -\sum_{i=1}^{|X|} P(i) \log P(i), \\ I(Y;C) &= H(Y) - H(Y|C), \\ NMI(Y,C) &= \frac{H(Y) + H(C)}{2} \times I(Y;C), \end{aligned} \tag{12}$$

for $NMI$, $Y$ represents the true labels of the data, and $C$ represents the predicted labels. $P(i)$ is the probability of random variable $X$ taking the value $i$, $H(Y)$ is the entropy of the true labels, and $H(Y|C)$ is the conditional entropy of the data given the clustering results $C$. Normalized Mutual Information maps the value of mutual information to the interval $[0, 1]$.

### 5.4 Experimental Settings

Among the aforementioned baseline algorithms, DeepWalk, GCN, and GAT are designed for homogeneous graphs. In homogeneous networks, they utilize the network's topology and node attributes to learn embeddings for nodes. On the other hand, Metapath2vec and HAN are developed for heterogeneous graphs, employing methods based on meta-paths. In this paper, we utilize an adaptive sampling strategy to sample the neighborhoods of target nodes instead of using meta-paths.

In our experiments, we set the learning rate to 0.01 to ensure the stability of the model training process. The dimension of node feature embeddings is set to 128. We use the following versions of software packages: **PyTorch** 1.2.0 as our deep learning framework, providing powerful tensor computation and deep neural network support. We use **DGL** 0.3.1 to handle data with graph structures. We utilize **NetworkX** 2.3 to create, manipulate, and study the structure and functions of complex networks. We apply **scikit-learn** 0.21.3 to train and evaluate machine learning

models. We use **NumPy** 1.17.2 for efficient multidimensional array operations. Finally, we use **SciPy** 1.3.1 for scientific and technical computing. These software packages and their versions were chosen based on our experimental requirements.

### 5.5 Experimental Results and Analysis

In this section, we conduct a series of simulation experiments using academic network datasets. The main purpose of these experiments is to evaluate the performance of our proposed ASGNN algorithm. This is achieved through a detailed comparison of overall experimental results and analysis of experimental parameter settings. Our experiments and subsequent discussions not only provide evidence of the effectiveness of the algorithm but also contribute to a deeper understanding of the importance of graph enhancement for enhancing model performance.

1. Analysis of the Overall Experimental Results.

   First, we conduct experiments and compare our proposed ASGNN model with some common methods, as shown in Table 2. Our ASGNN model achieves good experimental results in terms of *NMI* and *ARI* on the DBLP and ACM datasets. Additionally, to visualize the experimental results more clearly, we present them graphically in Figure 4 to better illustrate the effectiveness of our method. We make the following observations:

   (a) It can be observed that our proposed method achieves optimal *NMI* and *ARI* scores on both datasets. Because we employed an advanced approach to complete node attributes: first, by utilizing random walk techniques to capture the network's topology, and then applying an adaptive sampling algorithm to select the most influential neighboring node set, rather than relying on traditional metapath techniques.
   (b) The HAN model achieved suboptimal performance on both datasets. This is because, in contrast to Metapath2vec, which does not differentiate the importance of metapaths, HAN distinguishes the importance of metapaths, allowing it to achieve better results.
   (c) The GCN and GAT methods designed for homogeneous graphs do not perform well on heterogeneous academic network datasets.

   These experimental results indicate that our proposed ASGNN model can effectively enhance heterogeneous academic networks and achieve good performance by leveraging their rich information and structure. This underscores the effectiveness of our proposed model.

2. Ablation Experiment.

   To assess the impact of various key components of the ASGNN model on overall performance, we conducted an ablation study by sequentially removing individual components from the DBLP and ACM datasets. The experimental results are shown in Figure 5.

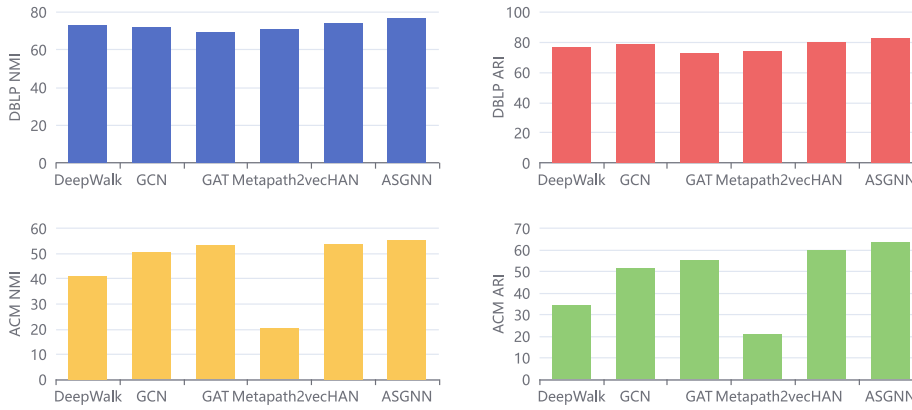



Figure 4. NMI and ARI comparison on the DBLP and ACM datasets

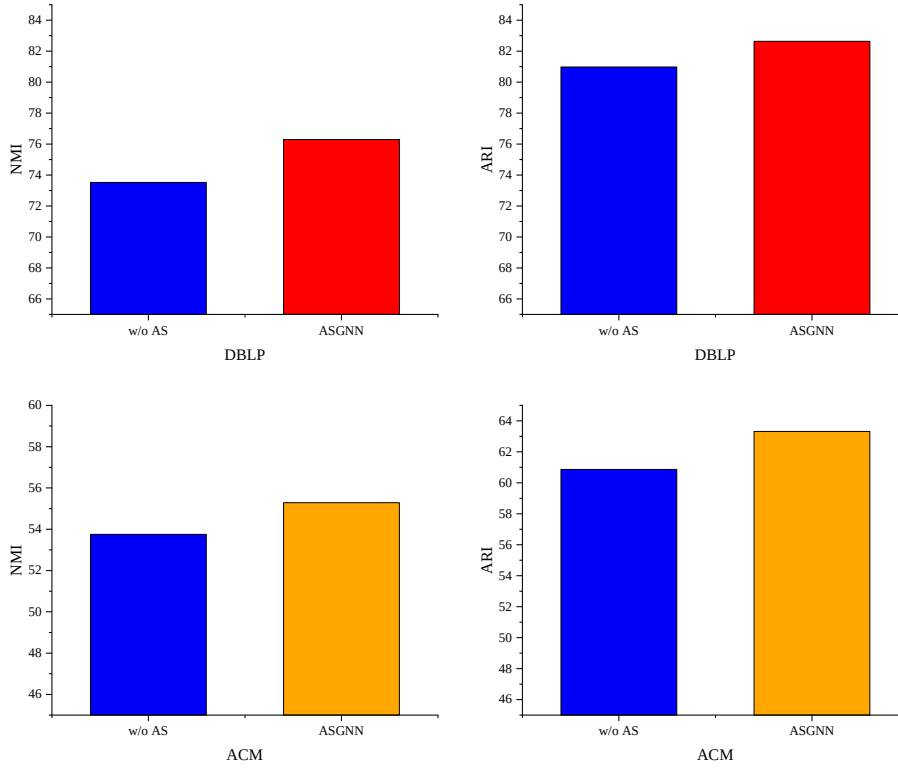



Figure 5. Performance of ASGNN and its variants

| **Metrics** | **DBLP** | | **ACM** | |
|---|---|---|---|---|
| | **NMI** | **ARI** | **NMI** | **ARI** |
| DeepWalk | 72.68 | 76.54 | 41.06 | 34.21 |
| GCN | 71.85 | 78.62 | 50.34 | 51.28 |
| GAT | 68.94 | 72.25 | 53.14 | 54.96 |
| Metapath2vec | 70.68 | 73.58 | 20.24 | 20.65 |
| HAN | 73.95 | 79.69 | 53.64 | 59.68 |
| ASGNN | **76.29** | **82.64** | **55.29** | **63.32** |

Table 2. Comparison of different metrics on two public networks

We evaluated the ASGNN model by selecting a quantitative number of neighbors for adaptive sampling. When the module was removed, denoted as "w/o QS", the results showed a significant decrease in model performance. This highlights the effectiveness of our adaptive sampling module in the model.

### 5.6 Visualization

For a more intuitive comparison among various methods including ours, we conducted visualization experiments aiming to reduce the dimensionality of node embeddings for better observation. Specifically, we employed PCA [46] and t-SNE [47] techniques to project the learned embeddings into a two-dimensional space. Taking the DBLP dataset as an example, we colored the data points according to their paper categories. The results are depicted in Figure 6 and Figure 7, where t-SNE demonstrates superior dimensionality reduction compared to PCA.

Figures 6 and 7 both reflect that GCN and GAT, designed for homogeneous graphs, perform poorly, resulting in authors from different research fields being mixed together. HAN performs much better than the above models based on homogeneous GNNs, but its boundaries are still blurry. Our proposed model enhances node attributes and then enhances the topological structure. By aggregating the enhanced topology with the node attribute information, better results can be achieved. Different types of papers are aggregated into different clusters, and the classification boundaries are very clear.

## 6 CONCLUSIONS

This paper presents an advanced method for heterogeneous GNN, which is mainly divided into two modules. One is the adaptive sampling module, which samples neighborhood nodes strongly correlated with each target node in an adaptive manner. By completing the missing information in heterogeneous academic networks, the network contains richer and more complete semantic information. The second module is the topological structure enhancement module. Some network enhancement models focus too much on node attributes and neglect the topological structure of the network. We compare the correlation between nodes of a certain type and

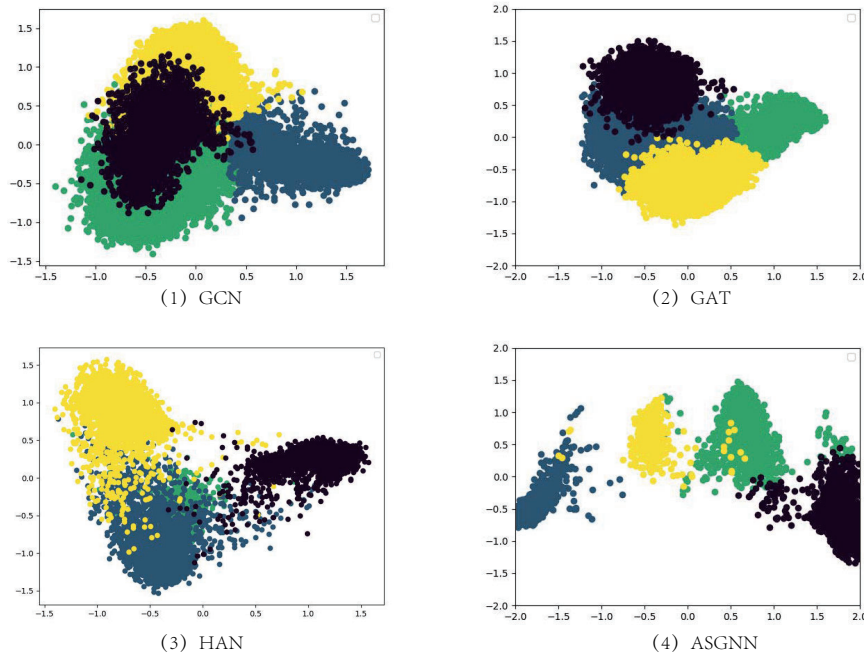



Figure 6. PCA visualization results

the target node with the average correlation score of nodes of that type induced by the target node's $n$-order subgraph to determine whether to add virtual edges, thereby enhancing the network topology. However, our model has some limitations. For example, the relationships between nodes may fluctuate over time, and its generalization ability in dynamic network datasets may be limited.

Our future work involves addressing multimodal information integration by considering the introduction of more types of node information, such as text, images, time, etc., to build a richer academic network. Integrating information from different modalities into the model to enhance the model's ability to handle diversified information. Additionally, considering the introduction of adversarial training mechanisms, such as GANs, to improve the model's robustness to outlier nodes or noise, enhancing the stability and generalization performance of the model. Furthermore, considering the evolution of node attributes and relationships in dynamic networks, we will optimize the model to adapt to the inherent dynamics of dynamic networks, enhancing the model's robustness and adaptability.

## Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2022YFB4501704, in part by the Na-

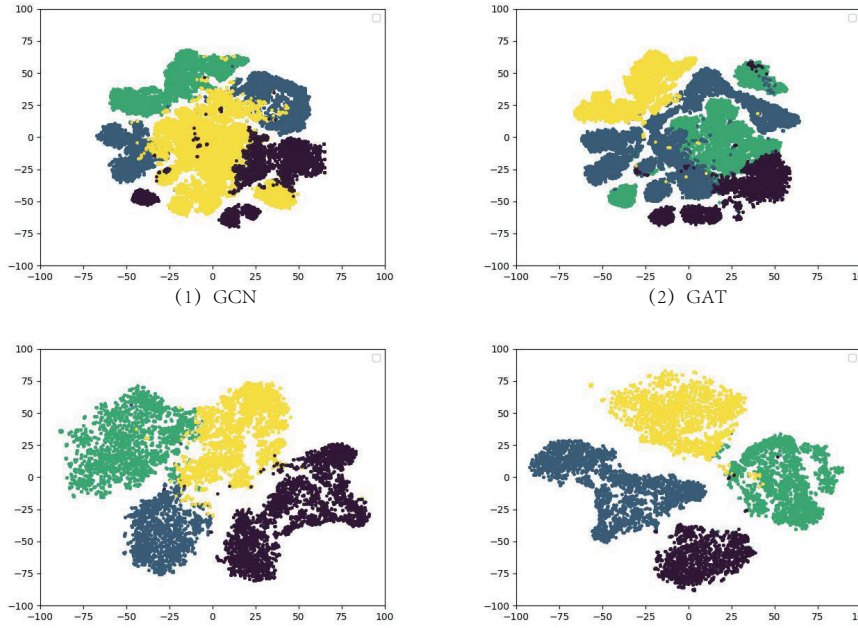



Figure 7. t-SNE visualization results

tional Natural Science Foundation of China under Grants No. 62302308, 62372300, 61702333 and U2142206, and in part by the Shanghai Sailing Program under Grant No. 21YF1432900.

## REFERENCES

[1] Wang, W.—Liang, J.—Yang, J.: Research on the Identification Method of the Topic Hierarchical Structure in a Domain Based on the Citation Network. Library and Information Service, Vol. 66, 2022, No. 17, pp. 81–92, doi: 10.13266/j.issn.0252-3116.2022.17.008 (in Chinese).

[2] Zhao, F.—Zhang, Y.—Lu, J.—Shai, O.: Measuring Academic Influence Using Heterogeneous Author-Citation Networks. Scientometrics, Vol. 118, 2019, No. 3, pp. 1119–1140, doi: 10.1007/s11192-019-03010-5.

[3] Wu, H.—Sun, Y.: On Status Quo of Citation Network Research and the Overview on Its Development. Computer Applications and Software, Vol. 29, 2012, No. 2, pp. 164–168, doi: 10.3969/j.issn.1000-386X.2012.02.048 (in Chinese).

[4] Hadhiatma, A.—Azhari, A.—Suyanto, Y.: A Scientific Paper Recommendation Framework Based on Multi-Topic Communities and Modified PageRank. IEEE Access, Vol. 11, 2023, pp. 25303–25317, doi: 10.1109/ACCESS.2023.3251189.

[5] Jin, T.—Wu, Q.—Ou, X.—Yu, J.: Community Detection and Co-Author Recommendation in Co-Author Networks. International Journal of Machine Learning and Cybernetics, Vol. 12, 2021, No. 2, pp. 597–609, doi: 10.1007/s13042-020-01190-8.

[6] Chen, J.—Ban, Z.: Academic Paper Recommendation Based on Clustering and Pattern Matching. In: Knight, K., Zhang, C., Holmes, G., Zhang, M. L. (Eds.): Artificial Intelligence (ICAI 2019). Springer, Singapore, Communications in Computer and Information Science, Vol. 1001, 2019, pp. 171–182, doi: 10.1007/978-981-32-9298-7_14.

[7] Gao, T.—Zhang, Y.—Wang, S.—Yang, Y.—Pan, R.: Community Detection for Statistical Citation Network by D-SCORE. Statistics and Its Interface, Vol. 14, 2021, No. 3, pp. 279–294, doi: 10.4310/20-SII636.

[8] Yudhoatmojo, S. B.—Samuar, M. A.: Community Detection on Citation Network of DBLP Data Sample Set Using LinkRank Algorithm. Procedia Computer Science, Vol. 124, 2017, pp. 29–37, doi: 10.1016/j.procs.2017.12.126.

[9] Sun, Y.—Han, J.—Zhao, P.—Yin, Z.—Cheng, H.—Wu, T.: RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '09), 2009, pp. 565–576, doi: 10.1145/1516360.1516426.

[10] Luo, L.—Fang, Y.—Cao, X.—Zhang, X.—Zhang, W.: CP-GNN: A Software for Community Detection in Heterogeneous Information Networks. Software Impacts, Vol. 10, 2021, Art. No. 100169, doi: 10.1016/j.simpa.2021.100169.

[11] Liang, X.—Ma, Y.—Cheng, G.—Fan, C.—Yang, Y.—Liu, Z.: Meta-Path-Based Heterogeneous Graph Neural Networks in Academic Network. International Journal of Machine Learning and Cybernetics, Vol. 13, 2022, No. 6, pp. 1553–1569, doi: 10.1007/s13042-021-01465-8.

[12] Liu, L.—Wang, S.: Meta-Path-Based Outlier Detection in Heterogeneous Information Network. Frontiers of Computer Science, Vol. 14, 2020, No. 2, pp. 388–403, doi: 10.1007/s11704-018-7289-4.

[13] Han, L.—Qin, J.—Xia, B.: Enhanced Social Recommendation Method Integrating Rating Bias Offsets. Electronics, Vol. 12, 2023, No. 18, Art. No. 3926, doi: 10.3390/electronics12183926.

[14] Wang, X.—Ji, H.—Shi, C.—Wang, B.—Ye, Y.—Cui, P.—Yu, P. S.: Heterogeneous Graph Attention Network. The World Wide Web Conference (WWW '19), 2019, pp. 2022–2032, doi: 10.1145/3308558.3313562.

[15] Luo, L.—Fang, Y.—Cao, X.—Zhang, X.—Zhang, W.: Detecting Communities from Heterogeneous Graphs: A Context Path-Based Graph Neural Network Model. Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21), 2021, pp. 1170–1180, doi: 10.1145/3459637.3482250.

[16] Zhang, C.—Song, D.—Huang, C.—Swami, A.—Chawla, N. V.: Heterogeneous Graph Neural Network. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19), 2019, pp. 793–803, doi: 10.1145/3292500.3330961.

[17] Zhang, S.—Jin, Y.—Liu, T.—Wang, Q.—Zhang, Z.—Zhao, S.—Shan, B.:

SS-GNN: A Simple-Structured Graph Neural Network for Affinity Prediction. ACS Omega, Vol. 8, 2023, No. 25, pp. 22496–22507, doi: 10.1021/acsomega.3c00085.

[18] Zhang, W.—Yin, Z.—Sheng, Z.—Li, Y.—Ouyang, W.—Li, X.—Tao, Y.—Yang, Z.—Cui, B.: Graph Attention Multi-Layer Perceptron. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), 2022, pp. 4560–4570, doi: 10.1145/3534678.3539121.

[19] Zeng, D.—Liu, W.—Chen, W.—Zhou, L.—Zhang, M.—Qu, H.: Substructure Aware Graph Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, No. 9, pp. 11129–11137, doi: 10.1609/aaai.v37i9.26318.

[20] Fu, X.—Zhang, J.—Meng, Z.—King, I.: MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding. Proceedings of the Web Conference 2020 (WWW '20), 2020, pp. 2331–2341, doi: 10.1145/3366423.3380297.

[21] Xu, G.—Meng, Y.—Qiu, X.—Yu, Z.—Wu, X.: Sentiment Analysis of Comment Texts Based on BiLSTM. IEEE Access, Vol. 7, 2019, pp. 51522–51532, doi: 10.1109/ACCESS.2019.2909919.

[22] Kaibiao, L.—Chen, J.—Ruicong, C.—Fan, Y.—Yang, Z.—Min, L.—Ping, L.: Adaptive Neighbor Graph Aggregated Graph Attention Network for Heterogeneous Graph Embedding. ACM Transactions on Knowledge Discovery from Data, Vol. 18, 2023, No. 1, Art. No. 29, doi: 10.1145/3616377.

[23] Zhu, G.—Zhu, Z.—Wang, W.—Xu, Z.—Yuan, C.—Huang, Y.: AutoAC: Towards Automated Attribute Completion for Heterogeneous Graph Neural Network. 2023 IEEE 39th International Conference on Data Engineering (ICDE), 2023, pp. 2808–2821, doi: 10.1109/ICDE55515.2023.00215.

[24] Zhang, S. S.—Chen, A. X.—Zhang, Z. B.: Fine-Tuned Heterogeneous Graph Convolutional Network Embedding. 2023 International Conference on Machine Learning and Cybernetics (ICMLC), IEEE, 2023, pp. 358–364, doi: 10.1109/ICMLC58545.2023.10327935.

[25] Zhao, Q.—Miao, Y.—An, D.—Lian, J.—Li, M.: HGNN-QSSA: Heterogeneous Graph Neural Networks with Quantitative Sampling and Structure-Aware Attention. IEEE Access, Vol. 12, 2024, pp. 25512–25524, doi: 10.1109/ACCESS.2024.3366231.

[26] He, D.—Liang, C.—Huo, C.—Feng, Z.—Jin, D.—Yang, L.—Zhang, W.: Analyzing Heterogeneous Networks with Missing Attributes by Unsupervised Contrastive Learning. IEEE Transactions on Neural Networks and Learning Systems, Vol. 35, 2022, No. 4, pp. 4438–4450, doi: 10.1109/TNNLS.2022.3149997.

[27] Jin, D.—Huo, C.—Liang, C.—Yang, L.: Heterogeneous Graph Neural Network via Attribute Completion. Proceedings of the Web Conference 2021 (WWW '21), 2021, pp. 391–400, doi: 10.1145/3442381.3449914.

[28] Fu, N.—Zhao, Q.—Miao, Y.—Zhang, B.—Wang, D.: Representation Learning Method of Graph Convolutional Network Based on Structure Enhancement. Computing and Informatics, Vol. 41, 2022, No. 6, pp. 1563–1588, doi: 10.31577/cai_2022_6_1563.

[29] Li, C.—Yan, Y.—Fu, J.—Zhao, Z.—Zeng, Q.: HetReGAT-FC: Heterogeneous Residual Graph Attention Network via Feature Completion. Information Sciences, Vol. 632, 2023, pp. 424–438, doi: 10.1016/j.ins.2023.03.034.

[30] Chen, H.—Li, Y.—Sun, X.—Xu, G.—Yin, H.: Temporal Meta-Path Guided Explainable Recommendation. Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21), 2021, pp. 1056–1064, doi: 10.1145/3437963.3441762.

[31] Wang, X.—Lu, Y.—Shi, C.—Wang, R.—Cui, P.—Mou, S.: Dynamic Heterogeneous Information Network Embedding with Meta-Path Based Proximity. IEEE Transactions on Knowledge and Data Engineering, Vol. 34, 2022, No. 3, pp. 1117–1132, doi: 10.1109/TKDE.2020.2993870.

[32] Zhang, J.—Yu, P. S.—Zhou, Z. H.: Meta-Path Based Multi-Network Collective Link Prediction. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14), 2014, pp. 1286–1295, doi: 10.1145/2623330.2623645.

[33] Hu, B.—Fang, Y.—Shi, C.: Adversarial Learning on Heterogeneous Information Networks. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19), 2019, pp. 120–129, doi: 10.1145/3292500.3330970.

[34] Wang, Z.—Yu, D.—Li, Q.—Shen, S.—Yao, S.: SR-HGN: Semantic- and Relation-Aware Heterogeneous Graph Neural Network. Expert Systems with Applications, Vol. 224, 2023, Art. No. 119982, doi: 10.1016/j.eswa.2023.119982.

[35] Zhong, H.—Wang, M.—Zhang, X.: HeMGNN: Heterogeneous Network Embedding Based on a Mixed Graph Neural Network. Electronics, Vol. 12, 2023, No. 9, Art. No. 2124, doi: 10.3390/electronics12092124.

[36] Hu, G.—Pang, J.: Relation-Aware Weighted Embedding for Heterogeneous Graphs. Information Technology and Control, Vol. 52, 2023, No. 1, pp. 199–214, doi: 10.5755/j01.itc.52.1.32390.

[37] Zhang, C.—Li, K.—Wang, S.—Zhou, B.—Wang, L.—Sun, F.: Learning Heterogeneous Graph Embedding with Metapath-Based Aggregation for Link Prediction. Mathematics, Vol. 11, 2023, No. 3, Art. No. 578, doi: 10.3390/math11030578.

[38] Lou, X.—Liu, G.—Li, J.: Heterogeneous Graph Neural Network with Graph-Data Augmentation and Adaptive Denoising. Applied Intelligence, Vol. 54, 2024, No. 5, pp. 4411–4424, doi: 10.1007/s10489-024-05363-8.

[39] Liu, J.—Song, L.—Gao, L.—Shang, X.: MMAN: Metapath Based Multi-Level Graph Attention Networks for Heterogeneous Network Embedding (Student Abstract). Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, No. 11, pp. 13005–13006, doi: 10.1609/aaai.v36i11.21639.

[40] Wu, H.—James, R. G.—D'Souza, R. M.: Correlated Structural Evolution Within Multiplex Networks. Journal of Complex Networks, Vol. 8, 2020, No. 2, Art. No. cnaa014, doi: 10.1093/comnet/cnaa014.

[41] Perozzi, B.—Al-Rfou, R.—Skiena, S.: DeepWalk: Online Learning of Social Representations. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14), 2014, pp. 701–710, doi: 10.1145/2623330.2623732.

[42] Kipf, T. N.—Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. 2017, doi: 10.48550/arXiv.1609.02907.

[43] Veličković, P.—Cucurull, G.—Casanova, A.—Romero, A.—Liò, P.—Bengio, Y.: Graph Attention Networks. International Conference on Learning Representations (ICLR 2018), 2018, doi: 10.48550/arXiv.1710.10903.

[44] Dong, Y.—Chawla, N. V.—Swami, A.: metapath2vec: Scalable Representation Learning for Heterogeneous Networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17), 2017, pp. 135–144, doi: 10.1145/3097983.3098036.

[45] Strehl, A.—Ghosh, J.: Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research, Vol. 3, 2002, pp. 583–617.

[46] Martinez, A. M.—Kak, A. C.: PCA Versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, 2001, No. 2, pp. 228–233, doi: 10.1109/34.908974.

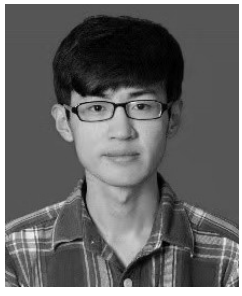
[47] van der Maaten, L.—Hinton, G.: Visualizing Data Using t-SNE. Journal of Machine Learning Research, Vol. 9, 2008, No. 11, pp. 2579–2605.

**Qin Zhao** received his Ph.D. degree from the Department of Computer Science and Technology, Tongji University, Shanghai, China, in 2016. Currently, he is an Associate Professor with the Department of Computer Science and Technology, Shanghai Normal University, Shanghai, China. He also serves as the Deputy Director of Shanghai Engineering Research Center of Intelligent Education and Big Data. His research interests include new-generation artificial intelligence technologies, social network analysis, data mining, and intelligence in scientific computing. He is a senior member of IEEE, and a distinguished member of the China Computer Federation (CCF). He has published over 50 papers in premier international journals and conferences, including IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE Transactions on Computational Social Systems, International Joint Conference on Artificial Intelligence, etc.

**Guojun Yang** received his B.Eng. degree in software engineering from the College of Information Engineering, Fuyang Normal University, Fuyang, China, in 2022. He is currently pursuing his M.Sc. degree in the Department of Computer Science and Technology, Shanghai Normal University, China. His research interests include community detection and social recommendation.

**Yaru Miao** received her B.Eng. degree from the Department of Computer, Shanghai Normal University in 2021. She is currently pursuing her M.Sc. degree in the Department of Computer Science and Technology at Shanghai Normal University, China. Her research interests include community detection, natural language processing and sentiment analysis.

**Jie Lian** received her Ph.D. degree from Towson University, Towson, MD, USA, in 2017. She is currently an Associate Professor with the Department of Computer Science and Technology, Shanghai Normal University, where she has been a Faculty Member, since 2017. Her research interests include spatio-temporal data mining, deep learning, and big data, ranging from theory to design to implementation. She received the Shanghai Sailing Talent Program, in 2019.



**Hongda Qi** received his Ph.D. degree in computer science and technology from Tongji University, Shanghai, China, in 2023. He is currently a Lecturer with the Department of Information, Mechanical and Electrical Engineering, Shanghai Normal University. His research interests include Petri net theory and machine learning.

**Zuliang Kou** received his dual Bachelor's degrees in computer science and technology, and business administration from Wuhan Institute of Science and Technology in 2006. In 2019, he was certified as a Senior Engineer in Computer Software and Hardware Development in Shanghai. In 2021, he graduated with a Master's degree in financial and risk management from the University of Alberta, Canada. His main research interests are the application of artificial intelligence in financial technology and educational practices.