Deep Generation of Heterogeneous Networks

Chen Ling, Carl Yang, and Liang Zhao*
Department of Computer Science, Emory University, USA
{chen.ling, j.carlyang, liang.zhao}@emory.edu
*Corresponding Author

Abstract—Heterogeneous graphs are ubiquitous data structures that can inherently capture multi-type and multi-modal interactions between objects. In recent years, research on encoding heterogeneous graph into latent representations have enjoyed a rapid increase. However, its reverse process, namely how to construct heterogeneous graphs from underlying representations and distributions have not been well explored due to several challenges in 1) modeling the local heterogeneous semantic distribution; 2) preserving the graph-structured distributions over the local semantics; and 3) characterizing the global heterogeneous graph distributions. To address these challenges, we propose a novel framework for heterogeneous graph generation (HGEN) that jointly captures the semantic, structural, and global distributions of heterogeneous graphs. Specifically, we propose a heterogeneous walk generator that hierarchically generates meta-paths and their path instances. In addition, a novel heterogeneous graph assembler is developed that can sample and combine the generated meta-path instances (e.g., walks) into heterogeneous graphs in a stratified manner. Theoretical analysis on the preservation of heterogeneous graph patterns by the proposed generation process has been performed. Extensive experiments[1] on multiple real-world and synthetic heterogeneous graph datasets demonstrate the effectiveness of the proposed HGEN in generating realistic heterogeneous graphs.

Index Terms—Heterogeneous Graph, Graph Generation, Deep Generative Models

I. Introduction

As a ubiquitous data structure, the graph can model connections (i.e., edges) between individual objects (i.e., nodes). Tremendous efforts have been made to study various types of graph problems, resulting in a rich literature of related papers and methods [1], [2], [3], [4]. The study of graphs can be mainly categorized into two categories: 1) graph representation learning, which aims at encoding graph topological and semantic information into vector space [5]; and 2) graph generation, which reversely aims at constructing graph-structured data from low-dimensional space containing the graph generation rules or distribution [6]. In the past years, previous studies of graphs have been made mostly on homogeneous graphs, which are the graphs consist of nodes under the same type. However, as a generalization of the homogeneous graph, heterogeneous graphs are the graphs with multiple types of nodes which further result in multiple types of edges, such as citation networks [7] and social networks [8], [9]. Figure 1(b) shows a citation network with author, paper, venue, and term as nodes and "authorship", "containment" and "publishment" as edges. The local semantics based on certain combinations

of node types and edge types reflect the key patterns of heterogeneous graphs [10], [11]. Such local semantics are typically represented as *meta-path*, a sequence of node types and edge types. Meta-paths characterize the rich and diverse relations among nodes [11], [12]. For example, as shown in Figure 1(b), two authors can be connected via a meta-path since they both contribute to a paper, while two authors can alternatively be connected because their papers are accepted at the same venue.

Due to the recent advancement of various graph neural network models, plenty of works [13], [14], [15], [16], [17], [18], [19], [20], [21] have been proposed on studying heterogeneous graph representation learning and embedding in the past few years. These works have achieved significant progress in many downstream tasks (e.g., meta-relation detection [13], [14], heterogeneous node embedding learning [15], [16], and heterogeneous link prediction [22], [17]). Among all the heterogeneous graph-related research, there remains a paucity of study on the *heterogeneous graph generation*. It is self-evident that the advantages of generating realistic heterogeneous graphs are at least two-faceted: 1) generating high-quality heterogeneous graphs requires us to comprehensively capture the latent graph distribution, which can greatly enrich our understanding of the implicit properties of heterogeneous graphs; 2) generating heterogeneous graphs is useful in specific downstream applications (e.g., recommendation system [23], knowledge graph reasoning [22], and node proximity search [10]). Despite the importance of the heterogeneous graph generation, in the past decade, only one study [24] tries to generate random heterogeneous graphs, which is based on hand-crafted rules and fails to generate realistic heterogeneous graph as it cannot learn the real data distribution underlying the observed graphs. On the other hand, a surge of research efforts on deep generative models [6], [25], [26], [2], [27], [28] have been recently observed in the task of homogeneous graph generation. Through learning latent and complex dependencies directly from observed graphs, these deep graph generative models leverage different ways to learn and capture the underlying graph-structure distributions directly from the observed data without the need for hand-crafted rules. These approaches have been shown superiority in maintaining the structural properties in homogeneous graphs.

However, existing deep generative models designed for homogeneous graphs cannot be trivially adapted to heterogeneous graphs due to the following significant challenges: 1) *Difficulties in preserving heterogeneous semantic information.*
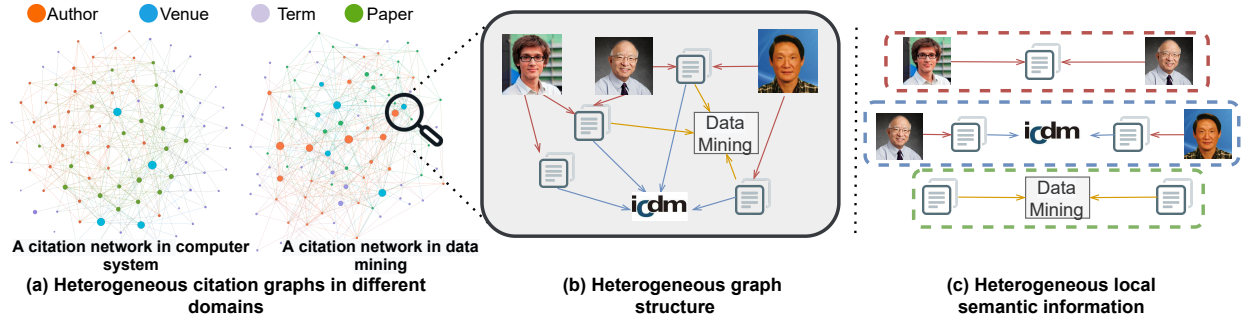
---

[1] https://github.com/lingchen0331/HGEN

Fig. 1: Examples of heterogeneous graph in academic field.

Current works for homogeneous graphs have been either using random walks as a tool to learn the graph topological distribution as learning the distribution of random walks ([26], [29]) or directly modeling an overall distribution of the edges ([30], [27]) over the homogeneous graphs. However, objects in heterogeneous graphs are inter-connected via various meta-paths as shown in Figure 1(c). As the complex local semantic information is carried by meta-paths, adapting current works to the heterogeneous graph scenario without any elaborations on meta-path would bring difficulties in learning and preserving the distribution of such complex semantic patterns spanning different graph entities (i.e., edges and nodes) in the newly generated heterogeneous graphs. 2) *Difficulties in preserving heterogeneous higher-order structural information.* In some cases, meta-paths may also fall short of expressing more intricate relationships among nodes in heterogeneous graphs. As marked in Figure 1(b), some common and symmetric higher-order structures spanning meta-paths will likely be observed repeatedly, which forms a triangle or orbit structure (e.g., one author writes two papers that are accepted by the same venue, and two papers of an author focus on the same research topic). The distributions of these higher-order graph structures are also hard to capture in heterogeneous graphs, which brings more challenges to effective heterogeneous graph generation. 3) *Difficulties in preserving heterogeneous global information.* Meta-paths are also well-recognized to play a fundamental role in preserving the global patterns of heterogeneous graphs [10]. For example, the ratio of different node types, and edge types, and their meta-paths are apparently different between the citation networks of computer system domain and data mining domain, as shown in Figure 1(a). It is important to preserve the global distribution of meta-path patterns during heterogeneous graph generation, which is again extremely difficult as it is entangled with the preservation of node type ratios, edge type ratios, and graph topological patterns.

In coping with these challenges, we introduce an end-to-end graph generative framework, namely <u>H</u>eterogeneous <u>G</u>raph <u>Gen</u>eration (HGEN), whose goal is to generate novel heterogeneous graphs by preserving all the complex local semantic, higher-order structural, and global properties through directly modeling the distribution of meta-paths in observed heterogeneous graphs. Particularly, to deal with the first challenge of capturing the complicated local semantics, we propose to learn a joint distribution of the random walks and the asso-

ciated meta-paths from the observed heterogeneous graphs. On top of that, for challenge two, we encode heterogeneous higher-order structural information into nodes via embedding learning and use it to guide the generation of meta-paths and random walks that form different high-order heterogeneous structures. To tackle the third challenge, we develop a novel heterogeneous graph assembly method, which is theoretically proved to preserve the global heterogeneous graph patterns in node types, edge types, and meta-paths. We conclude our major contributions as follows:

- **Problem.** We not only formulate a new paradigm of heterogeneous graph generation but also identify and resolve its unique challenges in preserving various heterogeneous graph properties.
- **Framework.** We propose an end-to-end generative framework for heterogeneous graph generation. The proposed framework can effectively learn the underlying distribution of heterogeneous graphs. It generates heterogeneous graphs with ensuring the preservation of various heterogeneous graph properties.
- **Evaluation.** We conduct extensive experiments on both synthetic and real-world heterogeneous graphs. Compared with state-of-the-art baselines, HGEN achieves competitive results in preserving most of the static graph properties. In addition, HGEN is shown to be capable of generating realistic heterogeneous graphs by preserving important meta-path information.

II. RELATED WORK

Heterogeneous Graph Mining. Compared to the commonly-adopted homogeneous graph, heterogeneous graph carries much richer semantic information and has therefore gained much attention in recent literature [31]. The concept of meta-paths in heterogeneous graph [12], [10] is one of the most important concepts proposed to capture numerous semantic relationships across multiple types of objects systematically. Since the introduction of heterogeneous graph, many innovative data mining tasks have spawned, including similarity search [10], object clustering [12], and heterogeneous node classification [15].

Heterogeneous Graph Representation Learning. In recent years, graph neural network (GNN) has achieved massive success in extensive applications [32], [3] due to its capability of effectively learning relationships and interactions on

non-Euclidean data. There exist plenty of attempts trying to adopt GNNs to learn with heterogeneous graphs, and almost all of them rely on employing meta-paths to model heterogeneous structures [20]. Specifically, proximity-preserving methods [13], [14], [17], [19] aim to capture heterogeneous network topological information via meta-path-constrained random walks. On the other line of approach, [18], [15], [16] try to aggregate information from heterogeneous neighbors via multiple layers of learnable projection functions. Throughout the study of heterogeneous graphs [10], [20], *meta-path* serves as the fundamental building block owing to its nonpareil ability to carry both graph topological and rich semantic information.

Graph Generation. Generative models for graphs have a rich history due to the wide range of applications in different domains, such as link prediction [26], [27], protein structure analysis [33], and information diffusion analysis in social networks [34]. Traditional graph generation methods (e.g., random graphs, stochastic block models, and Bayesian network models) fail to model complex dependencies in our real-world scenarios. In addition, they cannot effectively preserve the statistical properties of the observed graphs. In the last few years, there has been a surge in research focusing on deep graph generation. According to [6], the current deep graph generation can be divided into two categories: sequential-based and one-shot-based. For sequential-based graph generation methods [2], [26], [1], they autoregressively generate the nodes and edges with the LSTM model. However, the sequential-based generation is limited in following a fixed node/edge permutation order, which greatly loses the generation flexibility and model scalability. On the other hand, one-shot-based generation methods [27], [33], [26], [35], [28] try to build a probabilistic graph model based on the matrix representation that can generate graph topology as well as node/edge attributes in a one-shot, but most of them cannot easily be applied in large graphs due to the large time complexity. Finally, multi-attributed graph generation [2], [25], [36], [37] aims at generating homogeneous graphs by preserving node/edge attributes. Instead, the key patterns of heterogeneous graphs are the higher-order local semantics reflected by the combinatorial of the types of nodes and edges, which cannot be captured by methods for homogeneous graphs.

III. PROBLEM FORMULATION

A heterogeneous graph [31], [20] is a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with multiple types of objects and relations. $\mathcal{V}$ is the set of objects (i.e., nodes), where each node $v_i \in \mathcal{V}$ is associated with a node type $o = \phi(v_i)$. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, where each edge $e_{ij} \in \mathcal{E}$ is associated with a relation type $l = \psi(e_{ij})$.

In the study of heterogeneous graphs, the concepts of meta-paths are widely considered as cornerstones and adopted to systematically capture numerous semantic relationships across multiple types of objects, which are defined as a path over the graph [12], [20]. Hence meta-paths are indispensable to be considered as basic units for heterogeneous graph generation. Concretely, a meta-path $\mathbf{o}$ is defined as a sequence of object

types and edge types $\mathbf{o} = \big((o_1, o_2, ..., o_n), (l_1, l_2, ..., l_{n-1})\big) = o_1 \xrightarrow{l_1} o_2 \xrightarrow{l_2} ... \xrightarrow{l_{n-1}} o_n$, where each o_i and l_j are node type and edge type in the sequence, respectively. Each meta-path captures the rich semantic information between its two ends o_1 and o_n. In heterogeneous graphs, the local semantic information is carried on each of walks $\mathbf{v} = (v_0, v_1, ..., v_n)$ and its associated meta-path $\mathbf{o}$. We again take Figure 1(c) as an example, there exist two meta-paths between papers: (*Paper, Author, Paper*) and (*Paper, Venue, Paper*). The utilization of different meta-paths allow the heterogeneous graph to contain rich topological and semantics among diverse objects, which has been shown beneficial to many real-world graph mining applications [20], [15], [16].

With the preliminary notion of the heterogeneous graph, we formalize the heterogeneous graph generation problem as follows:

Problem 1 (**Heterogeneous Graph Generation**). *The goal of the heterogeneous graph generation is to learn a distribution $p_{data}(\mathcal{G})$ from the observed heterogeneous graphs such that a new graph $\hat{\mathcal{G}}$ can be obtained by sampling $\hat{\mathcal{G}} \sim p_{data}(\mathcal{G})$.*

Challenge 1 (**Difficulties in modeling the complex local semantic information.**). *Although the existence of meta-paths allows heterogeneous graph to characterize the combinatorial of node types and edge types, it is unclear how to model their distributions and generatively assemble them into heterogeneous graphs.*

Challenge 2 (**Difficulties in characterizing the heterogeneous structural patterns.**). *The local structural patterns in heterogeneous graphs are often expressed in higher-order proximity among the nodes and edges (e.g., triangles, orbits, and other higher-order structures). Such the local structure may fuse multiple walks under one or more meta-paths with richer semantic information, yet brings more difficulties in learning its distribution.*

Challenge 3 (**Difficulties in capturing heterogeneous global meta-path information.**). *Meta-paths indeed play a significant role in preserving the global patterns of heterogeneous graphs. In heterogeneous graph generation, it is important yet challenging to preserve the global distribution of meta-path patterns since the distribution of meta-path patterns often involves node type ratios, edge type ratios, and graph topological patterns.*

IV. HETEROGENEOUS GRAPH GENERATION

To address the above challenges, we propose a new heterogeneous graph generation framework, named HGEN. To address the first and second challenge, we propose a *heterogeneous walk generator* in Section IV-A to jointly learn the distribution of local walks and the associated meta-paths so that both heterogeneous topological and local semantic information can be well captured. To overcome the second challenge, we leverage the heterogeneous node embedding to make the generator be aware of any potential higher-order structures that each node may be involved with. Finally,
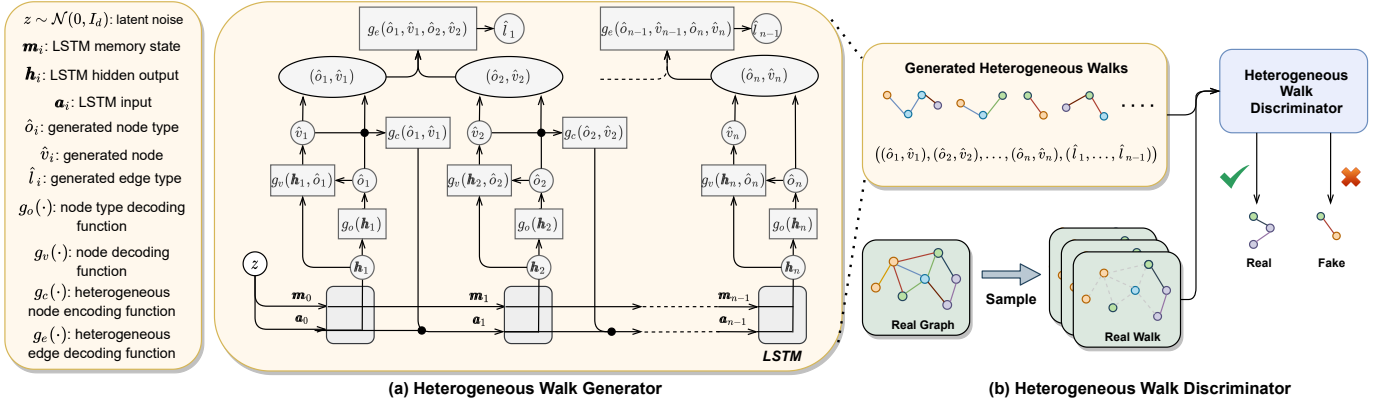
Fig. 2: The illustration of the heterogeneous walks generation in HGEN.

for the third challenge, we propose a novel *heterogeneous graph assembler* in Section IV-B, which can construct new heterogeneous graphs by capturing the global heterogeneous property, namely different meta-path ratios. We further prove that the global heterogeneous property can be well-preserved through our Theorem 1 introduced in Section IV-C.

A. Heterogeneous Walk Generator

In the observed graph $\mathcal{G}$, a heterogeneous walk is defined as a tuple that consists of two components: a walk $\mathbf{v}$ and an associated meta-path $\mathbf{o}$. The proposed heterogeneous walk generator G is defined as a probabilistic sequential learning model to generate synthetic heterogeneous walks: $(\hat{\mathbf{v}}, \hat{\mathbf{o}}) = \big((\hat{v}_1, \hat{v}_2, ..., \hat{v}_n), ((\hat{o}_1, \hat{o}_2, ..., \hat{o}_n), (\hat{l}_1, \hat{l}_2, ..., \hat{l}_{n-1}))\big)$, where the $\hat{\mathbf{v}}$ and $\hat{\mathbf{o}}$ are denoted as the generated walk and associated meta-path, respectively. We use $\hat{v}_i$, $\hat{o}_i$, and $\hat{l}_i$ to denote each of the generated node, node type, and edge type in $(\hat{\mathbf{v}}, \hat{\mathbf{o}})$, respectively. Figure 2(a) illustratively summarizes the whole generative process of each synthetic heterogeneous walk.

Heterogeneous Walk Generation. We model G as a sequential learning process based on a recurrent architecture, and each unit f_θ in the sequential model is parameterized by θ so that it can generate a node type $\hat{o}$ and a corresponding node $\hat{v}$ that belongs to this node type in a hierarchical manner. Precisely, the node type $\hat{o}$ is determined based on the previously generated sequence, and the node $\hat{v}$ is then coherently determined by the generated node type as well as the generated sequence. Both generated node type $\hat{o}$ and node $\hat{v}$ together provide information for the generation of the next node type and node instance.

Specifically, at each recurrent block (i.e., time step) t, f_θ produces two outputs $(\boldsymbol{m}_t, \boldsymbol{h}_t)$, where the $\boldsymbol{m}_t$ is the current memory state and the $\boldsymbol{h}_t$ is a latent probabilistic distribution (i.e., hidden output of f_θ) denoting the information carried from previous time steps. We first sample the node type $\hat{o}_t \sim g_o(\boldsymbol{h}_t)$ based on the probability distribution $\boldsymbol{h}_t$, where the $g_o(\cdot)$ is a node type decoding function. We then sample the node $\hat{v}_t$ by a node decoding function $\hat{v}_t \sim g_v(\boldsymbol{h}_t, \hat{o}_t)$ that takes $\boldsymbol{h}_t$ and $\hat{o}_t$ as inputs. Lastly, the generated node type $\hat{o}_t$ and node $\boldsymbol{h}_t$ are fused by a heterogeneous node encoding function $g_c(\hat{o}_t, \hat{v}_t)$, which then serves as the input of next recurrent block.

Heterogeneous Node Sampling. To overcome the second challenge, we cannot uniformly sample $\hat{v}_t$ based on the node type $\hat{o}_t$ because such a way may cause the neglection of (1) *node structural* distribution and (2) *node semantic* distribution. For example, we may observe an author always tends to cite a paper with high citation (namely, high node degree of this paper node). Then such distribution needs to be modeled with structural information. On the other hand, we may observe a data mining paper is unlikely to cite a computer system paper, and we may also need to characterize this tendency in the distribution. Both of the above distributions cannot be tackled by uniformly sampling. Therefore, to tackle this challenge, since latent node embedding could encode both topological and semantic information into the node, we propose to calculate a latent embedding $\tilde{v}_t$ of the next node v_t, then we select with a higher probability the closer embedding among all the embeddings that belong to node type $\hat{o}_t$ so that the next node v_t can be determined by the sampled embedding.

More specifically, we first calculate the latent node embedding $\tilde{v}_t$ based on the sampled node type $\hat{o}_t$ by a simple linear transformation. We then calculated the distance between $\tilde{v}_t$ and other node embedding $\tilde{v}_i^{(\hat{o}_t)}$, meaning any node $\tilde{v}_i$ belonging to the sampled node type $\hat{o}_t$. In this case, given a total number of k embeddings that belong to the type $\hat{o}_t$, the next node $\hat{v}_t$ can be sampled from a multinomial distribution:

$$\hat{v}_t \sim \text{Multi}(\tilde{v}_1^{(\hat{o}_t)}, \tilde{v}_2^{(\hat{o}_t)}, ..., \tilde{v}_k^{(\hat{o}_t)}; p_1, p_2, ..., p_k),$$

where each $p_i = -d(\tilde{v}_t, \tilde{v}_i^{(\hat{o}_t)})^2$ and $d(\cdot, \cdot)$ is a distance metric such as Euclidean distance. Note that the node embedding $\tilde{v}_i^{(\hat{o}_t)}$ can be obtained from a conventional heterogeneous node embedding technique such as [14].

In order to generate a variable-length heterogeneous walk, we incorporate a end-of-sequence token as an additional node type so that the heterogeneous walk generator stops when the sampled node type is the token at any steps. Therefore, the proposed generator is able to produce variable-length heterogeneous walks. Finally, the edge type l_t can be predicted by a simple edge decoding function $g_e(\hat{o}_t, \hat{v}_t, \hat{o}_{t-1}, \hat{v}_{t-1})$ that takes its two end nodes $\hat{v}_{t-1}$ and $\hat{v}_t$ as well as their node types $\hat{o}_{t-1}$ and $\hat{o}_t$ as inputs. In all, we summarize the overall

generative process as follows:

$$\boldsymbol{a}_0 = 0, \; \boldsymbol{m}_0 = f_0(\boldsymbol{z}), \; \boldsymbol{z} \sim \mathcal{N}(0,1)$$
$$\boldsymbol{a}_1 = g_c(\hat{o}_1, \hat{v}_1), \; \hat{v}_1 \sim g_v(\boldsymbol{h}_1, \hat{o}_1), \; \hat{o}_1 \sim g_o(\boldsymbol{h}_1), (\boldsymbol{m}_1, \boldsymbol{h}_1) = f_\theta(\boldsymbol{m}_0, \boldsymbol{a}_0)$$
$$\boldsymbol{a}_2 = g_c(\hat{o}_2, \hat{v}_2), \; \hat{v}_2 \sim g_v(\boldsymbol{h}_2, \hat{o}_2), \; \hat{o}_2 \sim g_o(\boldsymbol{h}_2), (\boldsymbol{m}_2, \boldsymbol{h}_2) = f_\theta(\boldsymbol{m}_1, \boldsymbol{a}_1)$$
$$\hat{l}_1 = g_e(\hat{o}_2, \hat{v}_2, \hat{o}_1, \hat{v}_1)$$
$$\dots$$
$$\hat{v}_n \sim g_v(\boldsymbol{h}_n, \hat{o}_n), \; \hat{o}_n \sim g_o(\boldsymbol{h}_n), \; (\boldsymbol{m}_n, \boldsymbol{h}_n) = f_\theta(\boldsymbol{m}_{n-1}, \boldsymbol{a}_{n-1})$$
$$\hat{l}_{n-1} = g_e(\hat{o}_n, \hat{v}_n, \hat{o}_{n-1}, \hat{v}_{n-1})$$

In this work, we utilize LSTM as the recurrent architecture, and f_θ becomes a single LSTM unit. To initialize the whole generative process, G takes a random noise $\boldsymbol{z}$ as input, which is drawn from a standard Gaussian distribution. Additionally, for the node type decoding function $g_o(\cdot)$, we apply the Gumbel-softmax trick [38] in $g_o(\cdot)$ to make the whole sampling differentiable. Finally, in most of the real-world scenarios, the edge type l_t can be determined by the types of its two end nodes $\hat{o}_t$ and $\hat{o}_{t-1}$ if there does not exist multi-typed relations between two node types. In this case, the heterogeneous walk generator can be simplified only to generate node sequences and associated node types.

B. Heterogeneous Generator Training and Utilization

In the following, we will introduce how to train the above-mentioned generator and how to use the heterogeneous walks generated by it to construct heterogeneous graphs. Concretely, we utilize a heterogeneous discriminator D to distinguish between real and fake heterogeneous walks, where the real heterogeneous walks are uniformly sampled from the observed graph. We then propose a heterogeneous graph assembler to construct new graphs based on the sampled heterogeneous walks. More details are presented as follows.

We first introduce the overall objective function of the Wasserstein heterogeneous GAN [39], which is written as:

$$\begin{aligned}
\mathcal{L}_{\text{HGEN}} = \max \; & \mathbb{E}_{(\boldsymbol{o}, \boldsymbol{v}) \sim p(\mathcal{G})}[D_o(\boldsymbol{o}) + D_v(\boldsymbol{v})] \\
& - \mathbb{E}_{\boldsymbol{z} \sim p(z)}[D_o(\hat{\boldsymbol{o}}) + D_v(\hat{\boldsymbol{v}})], s.t. \; G(z) = (\hat{\boldsymbol{o}}, \hat{\boldsymbol{v}}),
\end{aligned}$$
(1)

where $\boldsymbol{v}$ and $\boldsymbol{o}$ are the random walk and associated meta-path, respectively, directly sampled from the observed heterogeneous graph $\mathcal{G}$. They are the real data for training our heterogeneous walk generator G. Specifically, given an observed heterogeneous graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, we utilize random-walk-based method to uniformly sample a set of random walks $\{\boldsymbol{v}_1, \boldsymbol{v}_2, ...\}$, where each $\boldsymbol{v}_i$ is a node sequence s.t. $\boldsymbol{v}_i = (v_1, v_2, ..., v_n)$. In addition, we extract the meta-path information $\boldsymbol{o}_i = \big((o_1, o_2, ..., o_n), (l_1, l_2, l_{n-1})\big)$ from each $\boldsymbol{v}_i$.

The heterogeneous discriminator D in Equation (1) is designed as a parallel recurrent architecture in order to individually distinguish whether each component in the heterogeneous walks are valid or not. Specifically, at each recurrent block (i.e., each step) t, the discriminator D takes two inputs: the generated node type $\hat{o}_t$ and node index $\hat{v}_t$, each of which is fed into an individual recurrent unit. After processing both sequences, the discriminator returns a single score $D_v(\boldsymbol{v}) +$

$D_o(\boldsymbol{o})$ that represents the probability of the heterogeneous walk being real.

Heterogeneous Graph Assembler. To assemble a heterogeneous graph from the generated heterogeneous walks, we further propose a novel stratified heterogeneous edge sampling strategy to achieve the following steps: 1) it first samples a node $\hat{v}_i$ and its type $\hat{o}_i$ from all of the generated heterogeneous walks; 2) based on the node type $\hat{o}_i$, we then sample a meta-path that starts with $\hat{o}_i$; 3) we iteratively sample the next node $\hat{v}_{i+1}$ in the sampled meta-path if both of the node type $\hat{o}_{i+1}$ and edge type $\hat{l}_i$ fits the meta-path pattern.

More specifically, the generator G firstly produces a sufficient number of heterogeneous walks as shown in Figure 3(a). We then construct an symmetric adjacency matrix S with size $|\mathcal{V}| \times |\mathcal{V}|$ to record the count of edges observed from the sampled heterogeneous walks in each entry S_{ij}, where the $|\mathcal{V}|$ is the size of the node set. Next, we collect all of the meta-path patterns generated by the generated heterogeneous walks, as shown in Figure 3(b-c). For the first step of the stratified heterogeneous edge sampling, we sample the a node $\hat{v}_i$ and its type type $\hat{o}_i$ based on the node degree distribution $\frac{\sum_j S_{ij}}{|\mathcal{V}|}$. For the second step, among all the meta-paths $\{\boldsymbol{o}_1^{(f)}, \boldsymbol{o}_2^{(f)}, ...\}$ that start with the node type $\hat{o}_i$, we sample a meta-path $\boldsymbol{o}_i^{(f)}$ based on the probability $\frac{c(\boldsymbol{o}_i^{(f)})}{T^{\hat{o}_i}}$, where $T^{\hat{o}_i}$ is the total count of generated meta-paths that starts with node type $\hat{o}_i$ and $c(\boldsymbol{o}_i^{(f)})$ is the count of meta-path pattern $\boldsymbol{o}_i^{(f)}$. For the third step, by following this meta-path pattern $\boldsymbol{o}_r = (o_1, o_2, ..., o_n)$, we iteratively sample all the nodes whose node types are regulated by the the meta-path. Precisely, we sample the next node v_j by sampling all the neighbors of the current node v_i with the probability $p_{v_i v_j} = (S_{ij})/(\sum_s S_{is})$ such that all the nodes v_s belong to the specific node type o_j following the meta-path $\boldsymbol{o}_i^{(f)}$. The sampled node sequence $\boldsymbol{v}_r = (v_0, v_1, ...)$ is then added to the current under construction. We continue the stratified heterogeneous edge sampling strategy until the desired amount of edges is reached. The final assembled graph is visualized in Figure 3 (d).

Complexity Analysis. The computational complexity of HGEN is $O(W \cdot L)$, where W is the weights of a single LSTM unit, and L is the length of the generated heterogeneous walks. However, the length of our proposed heterogeneous walk is considerably small ($1 \leq L \leq 3$) while the walk length in other random-walk-based graph generative method [26] is (≥ 16). For auto-regressive graph generation models [3], [2], the time complexities are at least $O(|\mathcal{V}|^2 \cdot W)$, where $|\mathcal{V}|$ is the cardinality of the node set. They convert graph as a long sequence by performing a large number of breadth-first-search (BFS) enumerations for each graph. Additionally, HGEN also has linear complexity in graph assembly, it only needs to run the trained model T_s times to sample heterogeneous walks for constructing the score matrix S. To sum up, the overall complexity of HGEN can be reduced to $O(W + T_s)$, which makes our proposed model highly efficient for handling large graphs, since the overall process is not sensitive to the number
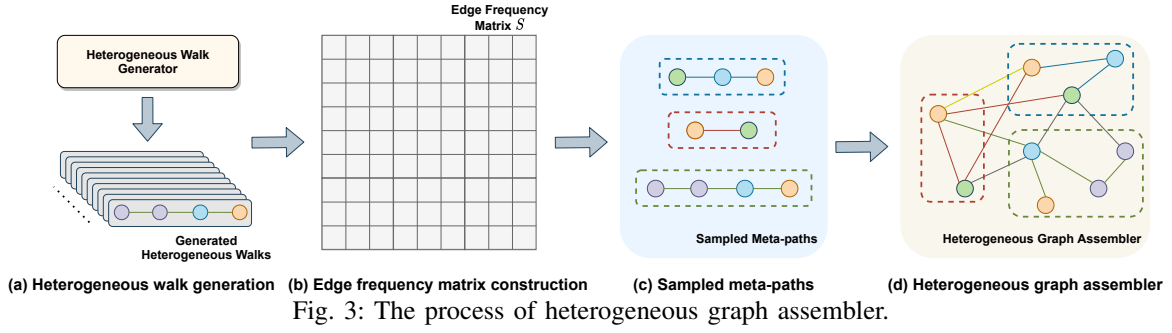
(a) Heterogeneous walk generation (b) Edge frequency matrix construction (c) Sampled meta-paths (d) Heterogeneous graph assembler

Fig. 3: The process of heterogeneous graph assembler.

(a) Observed heterogeneous graph **(b) Inaccurate generated graph**
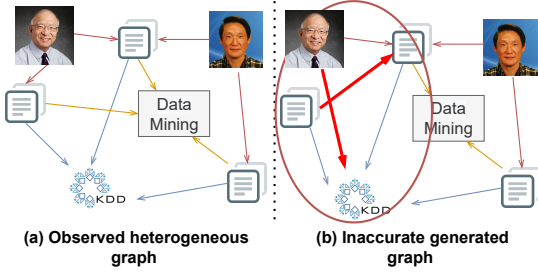
Fig. 4: Example of two heterogeneous graphs with different semantic information: the observed meta-path patterns are different, although the node and edge distribution are the same between two graphs. Specifically, since we do not observe a direct link between (author, venue) and (paper, paper) in the observed graph Figure 4(a). It is not accurate for the generated graph Figure 4(b) that generate such links.

of nodes at all.

C. Meta-path Information Preservation Analysis

As we discussed in Section III, it is significant to preserve the meta-path information in our generated graph. Taking Figure 4 as an example, although both graphs have exactly the same structure, they are still regarded as two different heterogeneous graphs since their meta-path distributions are different. Given the importance of the meta-path information in heterogeneous graph generation, we further show that our framework can successfully preserve this meta-path information as proved in **Theorem** 1.

Theorem 1. *The distribution of meta-path patterns $\overline{\mathcal{O}}^{(r)}$ of the generated heterogeneous graph equals the distribution of meta-path patterns $\overline{\mathcal{O}}$ in the observed heterogeneous graph, namely $p(\overline{\mathcal{O}}^{(r)}) = p(\overline{\mathcal{O}})$.*

Proof. We will prove that the ratio of the meta-path patterns can be preserved in three steps: 1) the ratio of different meta-path patterns can be preserved during the sampling procedure; 2) the ratio of generated meta-path patterns can be preserved during the generation procedure; 3) the meta-path patterns can be preserved during the graph assembling procedure.

Meta-path Ratio Preservation in Sampling. Let $\overline{\mathcal{O}} = (\overline{\mathbf{o}}_1, \overline{\mathbf{o}}_2, ...)$ be the collection of meta-paths obtained from the observed heterogeneous graph $\mathcal{G}$, each $\overline{\mathbf{o}}_i$ is a meta-path in one-hot format $\overline{\mathbf{o}}_i \in \{0,1\}^{1 \times R}$, where the R is the total number of different meta-path patterns. $\overline{\mathcal{O}}^{(\tau)} =$

$(\overline{\mathbf{o}}_1^{(\tau)}, \overline{\mathbf{o}}_2^{(\tau)}, ..., \overline{\mathbf{o}}_K^{(\tau)})$ is the sequence of sampled meta-paths with sampling size K, where each meta-path $\overline{\mathbf{o}}_j^{(\tau)} \in \{0,1\}^{1 \times R}$ is drawn independent and identically distributed ($i.i.d$) from $\overline{\mathcal{O}}$.

Suppose that $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_R]^T$ denotes the probability of each individual meta-path pattern in $\overline{\mathcal{O}}$, it is obvious that $\mathbb{E}[\overline{\mathbf{o}}_i|\boldsymbol{\mu}] = \sum_{\overline{\mathbf{o}}_i} p(\overline{\mathbf{o}}_i|\boldsymbol{\mu})\overline{\mathbf{o}}_i = [\mu_1, \mu_2, ..., \mu_R]^T = \boldsymbol{\mu}$. Now consider the total K observations $\overline{\mathcal{O}}^{(\tau)} = (\overline{\mathbf{o}}_1^{(\tau)}, \overline{\mathbf{o}}_2^{(\tau)}, ..., \overline{\mathbf{o}}_K^{(\tau)})$, the corresponding likelihood function takes the form:

$$p(\overline{\mathcal{O}}^{(\tau)}|\boldsymbol{\mu}) = \prod_i^R \prod_j^K \mu_j^{\overline{\mathbf{o}}_{ij}^{(\tau)}} = \prod_j^K \mu_j^{\sum_n \overline{\mathbf{o}}_{nj}^{(\tau)}} = \prod_j^K \mu_j^{m_j} \quad (2)$$

We see that the likelihood function depends on the K data points only through the R quantities: $m_j = \sum_n \overline{\mathbf{o}}_{nj}^{(\tau)}$. Since the number of observations of $\overline{\mathbf{o}}_j^{(\tau)}$ equals 1, we achieved sufficient statistics for this distribution. Therefore, $p(\overline{\mathcal{O}}^{(\tau)}) = p(\overline{\mathcal{O}})$ can be proved.

Meta-path Ratio Preservation in Generation. Since we have proved the meta-path ratio can be preserved during the sampling, the next step is to show that the distribution of generated meta-paths $p(\overline{\mathcal{O}}^{(g)})$ is equal to $p(\overline{\mathcal{O}}^{(\tau)})$. Proving $p(\overline{\mathcal{O}}^{(g)}) = p(\overline{\mathcal{O}}^{(\tau)})$ is equivalent to prove whether $p_{data} = p_g$ in the GAN setting. As being proved in the works of GANs and their variants [40], [39], it showed that the objective function of the generator G is equivalent to optimize the distribution distance between p_{data} and p_g if the discriminator D is optimal. Therefore, global optimality of $p_g = p_{data}$ can be achieved if both generator G and discriminator D have enough capability. Therefore, $p(\overline{\mathcal{O}}^{(g)}) = p(\overline{\mathcal{O}}^{(\tau)})$ if both G and D are optimal in our framework.

Meta-path Ratio Preservation in Assembling. Finally, we show that our graph assembling method can also preserve the meta-path ratio from the generated data $\overline{\mathcal{O}}^{(g)}$ such that $p(\overline{\mathcal{O}}^{(g)}) = p(\overline{\mathcal{O}}^{(r)})$. As we discussed in Section IV-B, the new graph $\hat{\mathcal{G}}$ is directly assembled by meta-paths $(\overline{\mathbf{o}}_1^{(g)}, \overline{\mathbf{o}}_2^{(g)}, ..., \overline{\mathbf{o}}_Q^{(g)})$ that are sampled $i.i.d$ from $\overline{\mathcal{O}}^{(g)}$ with sampling size Q, which is exactly the reverse procedure of Equation (2).

Therefore, if both generator G and discriminator D are optimal, the multinomial distribution $p(\overline{\mathcal{O}})$ of distinct meta-path patterns can be preserved in all three steps of our generation framework. $\square$

V. Experiment

In this section, we compare HGEN to the adaption of closest state-of-the-art baselines, demonstrating its effectiveness in generating realistic heterogeneous graphs in diverse settings. The code and datasets are made available[2].

A. Data

Synthetic Datasets. We synthesis random heterogeneous graphs of different sizes through the combination of N overlapping homogeneous graphs, where the overlap is accomplished by node sharing. We generate three random heterogeneous graphs (named as Syn_{100}, Syn_{200}, and Syn_{500}) with node size 100, 200, and 500, respectively. The number of node types in each of the synthetic heterogeneous graph is 3.

Real-world Datasets. We also employ three large-scale real-world heterogeneous graph datasets in our experiment.

- **PubMed**. This dataset consists of four classes of nodes: Gene (G), Disease (D), Chemical (C), and Species (S). We construct a sub-graph that relates to all Chemical nodes labeled in [20]. There are $1,565$ nodes and $13,532$ edges.
- **IMDB**. This movie-related heterogeneous graph is adopted from [15], which contains three node types: Director (D), Actor (A), Movie (M), and Genre (G). We construct a subgraph that contains all the movies with a score ≥ 7.5. This graph contains $1,653$ nodes and $4,267$ edges.
- **DBLP**. This heterogeneous graph adopted from [15] contains Paper (P), Author (A), Venue (V), and Term (T) as node types. We sample a subgraph that is related to five computer science venues: *KDD*, *WSDM*, *WWW*, *ICDM*, and *ICML*. There are $1,565$ nodes and $47,885$ edges.

B. Experiment Setting

In our experiment, we focus on meta-paths with length 1, 2, and 3 as they are the most common ones in heterogeneous graphs [10]. We sample 10 graphs from each of the trained models and report their average results and standard deviation in Table I. We randomly select 60% of the edges for training, and the remaining graph is used for testing.

Baselines. Since no baseline is available for the novel task of heterogeneous graph generation, we carefully adapt four state-of-the-art graph generation methods: NetGAN [26], GraphVAE [27], VGAE [30], and GraphRNN [2]. We utilize node type information as node features of the input graph in GraphVAE and VGAE. In addition, we modify NetGAN and GraphRNN to make them available to generate node types.

Evaluation Metrics. The evaluation of heterogeneous graph generation can be divided into three aspects. 1) *Graph Statistical Properties*: we focus on six typical statistics as widely used in [26], [35], [37] for measuring the structural similarity, including LCC (the size of the largest connected component), TC (Triangle count), Clustering Coef. (clustering coefficient); Powerlaw Coef. (power-law distribution of the node degree distribution), Assortativity, and Degree Distribution Dist. (Node degree distribution Maximum Mean

Discrepancy distance). 2) *Graph Novelty and Uniqueness.* Ideally, we would want the generated graphs to be diverse and similar, but not identical. To quantify this aspect, we check the uniqueness between the generated graphs by calculating their edit distances. Additionally, we calculate the EO Rate (edge overlapping rate) between the generated graphs and the testing graphs for measuring the novelty of the generated graphs. 3) *Meta-path Ratio Properties*: We measure the preservation of meta-path distribution in two metrics. Firstly, we measure the meta-path length ratio preservation. Secondly, under different meta-path lengths, we also measure the distribution of the frequent meta-path patterns.

C. Quantitative Analysis

Preservation of Graph Statistical Properties. We evaluate the performance of HGEN against all the baselines on the standard graph statistics, and the results are shown in Table I. Overall, HGEN achieves competitive performance constantly with very few exceptions on all metrics over both synthetic and real-world datasets. We report several observations from the table: 1) *Node-level similarity*: HGEN is the dominant performer in most node-level metrics. Although there are no significant differences in both Assortativity and Power-law Coef. among all the algorithms, HGEN rank top with very few exceptions in the node degree distribution distance with at least 40% improvement, which indicates that HGEN can effectively capture the degree distribution of all types of nodes through jointly learning both meta-path and random walk distribution. 2) *Graph level similarity*: HGEN still exceeds other baselines by effectively preserving the community distribution. Specifically, for all the datasets with rich local community information (e.g., PubMed and synthetic datasets), HGEN can utilize the heterogeneous node embedding for preserving the higher-order structural information in the generated heterogeneous walks, which leads to better performance in metrics like LCC, TC, and Clustering Coef.. However, in heterogeneous graphs with rare high-order structures, the performance of HGEN is comparatively less impressive. 3) As shown in Table I, the random-walk based method HGEN and NetGAN can generally achieve stable performance than one-shot based (e.g., VGAE and GraphVAE) and sequential-based (GraphRNN) generative models across all datasets. The reason is that random walk based methods learn the overall graph distribution by learning the distribution of its discrete random walks, which is not sensitive to various graph characteristics. 4) Table I also shows that VGAE cannot produce realistic graphs even though it achieves the best performance in some metrics, which is expected since the primary purpose of VGAE is learning node embeddings but not generating entire graphs. In addition, as the size of the graph increases, GraphRNN also fails to generate realistic graphs because of the weak scalability of auto-regressive models.

Graph Novelty and Uniqueness. The results of graph novelty and uniqueness are reported at the right two columns in Table I. Specifically, HGEN achieves a generally lower EO rate across all datasets, indicating that HGEN does not

Graphs	Models	LCC	TC	Clustering Coef.	Powerlaw Coef.	Assortativity	Degree Distribution Dist.	EO Rate	Uniqueness
Syn-100	GraphRNN	78.43 ± 2.23	16.62 ± 5.42	0.002 ± 0.01	1.611 ± 0.09	-0.153 ± 0.07	$2.19e{-}2 \pm 3.21e{-}3$	$37.21\% \pm 1.08\%$	$33.09\% \pm 7.06\%$
	NetGAN	80.12 ± 3.45	6.79 ± 1.76	0.001 ± 0.00	1.524 ± 0.21	-0.213 ± 0.09	$1.33e{-}2 \pm 6.46e{-}3$	$8.74\% \pm 0.82\%$	$\mathbf{94.03\% \pm 0.49\%}$
	GraphVAE	99.01 ± 0.00	224.81 ± 5.13	0.70 ± 0.04	4.579 ± 0.05	-0.73 ± 0.05	$3.71e{-}1 \pm 1.98e{-}2$	$11.5\% \pm 1.09\%$	$65.54\% \pm 2.98\%$
	VGAE	48.9 ± 4.63	63.7 ± 46.25	0.184 ± 0.06	$\mathbf{1.87 \pm 0.10}$	0.1 ± 0.03	$2.23e{-}1 \pm 6.08e{-}2$	$\mathbf{3.23\% \pm 0.09\%}$	$51.1\% \pm 3.04\%$
	HGEN	$\mathbf{81.13 \pm 2.42}$	$\mathbf{53.12 \pm 3.78}$	$\mathbf{0.079 \pm 0.01}$	1.782 ± 0.01	-0.114 ± 0.03	$\mathbf{8.79e{-}3 \pm 3.12e{-}3}$	$10.2\% \pm 0.17\%$	$92.97\% \pm 0.72\%$
	Real	85	36	0.072	1.832	-0.169	N/A	N/A	N/A
Syn-200	GraphRNN	132.76 ± 1.08	2.54 ± 0.77	0.001 ± 0.00	1.603 ± 0.01	-0.05 ± 0.01	$5.15e{-}2 \pm 3.07e{-}3$	$25.81\% \pm 2.65\%$	$27.72\% \pm 3.07\%$
	NetGAN	153 ± 1.56	2.24 ± 0.35	0.001 ± 0.00	1.579 ± 0.31	-0.008 ± 0.001	$6.43e{-}2 \pm 4.2e{-}3$	$11.32\% \pm 0.77\%$	$95.88\% \pm 3.19\%$
	GraphVAE	195.43 ± 1.12	51.32 ± 1.01	0.002 ± 0.001	5.377 ± 0.21	-0.75 ± 0.05	$5.38e{-}1 \pm 1.7e{-}2$	$\mathbf{1.78\% \pm 0.41\%}$	$64.37\% \pm 2.94\%$
	VGAE	86.2 ± 16.93	860.4 ± 185.9	0.23 ± 0.04	$\mathbf{1.787 \pm 0.08}$	0.2 ± 0.15	$8.53e{-}2 \pm 2.14e{-}2$	$3.74\% \pm 0.08\%$	$59.65\% \pm 1.46\%$
	HGEN	$\mathbf{158.5 \pm 2.64}$	$\mathbf{38.5 \pm 5.26}$	$\mathbf{0.043 \pm 0.01}$	1.732 ± 0.02	-0.065 ± 0.04	$\mathbf{2.25e{-}2 \pm 5.5e{-}3}$	$4.22\% \pm 0.67\%$	$\mathbf{96.31\% \pm 5.11\%}$
	Real	180	28	0.037	1.809	-0.089	N/A	N/A	N/A
Syn-500	GraphRNN	311.59 ± 2.14	11.53 ± 5.57	0.004 ± 0.001	1.862 ± 0.01	1.862 ± 0.002	$4.05e{-}2 \pm 1.1e{-}3$	$21.87\% \pm 0.86\%$	$29.54\% \pm 4.32\%$
	NetGAN	305.81 ± 14.28	$\mathbf{3 \pm 1.21}$	$\mathbf{0.001 \pm 0.001}$	1.812 ± 0.07	0.03 ± 0.12	$4.83e{-}2 \pm 7.4e{-}4$	$6.72\% \pm 0.13\%$	$93.98\% \pm 0.21\%$
	VGAE	97.0 ± 29.24	4346.2 ± 453.62	0.193 ± 0.02	1.77 ± 0.06	-0.022 ± 0.09	$2.22e{-}1 \pm 2.4e{-}2$	$5.46\% \pm 1.12\%$	$63.65\% \pm 3.1\%$
	HGEN	$\mathbf{347.88 \pm 7.63}$	74.88 ± 4.78	0.031 ± 0.01	$\mathbf{1.865 \pm 0.02}$	$\mathbf{-0.097 \pm 0.01}$	$\mathbf{2.81e{-}2 \pm 3.4e{-}3}$	$\mathbf{1.49\% \pm 0.11\%}$	$\mathbf{95.89\% \pm 1.18\%}$
	Real	417	8	6.5e−3	1.978	-0.12	N/A	N/A	N/A
PubMed	GraphRNN	1563.23 ± 32.46	1549.79 ± 33.62	0.01 ± 0.007	$\mathbf{1.753 \pm 0.04}$	-0.03 ± 0.01	$1.61e{-}1 \pm 3.71e{-}2$	$13.41\% \pm 1.24\%$	$54.62\% \pm 4.32\%$
	NetGAN	793.2 ± 41.5	18.3 ± 0.9	0.001 ± 0.00	1.47 ± 0.11	-0.12 ± 0.02	$6.69e{-}2 \pm 1.5e{-}3$	$4.32\% \pm 0.54\%$	$78.03\% \pm 0.19\%$
	VGAE	347.9 ± 7.03	$70,982.2 \pm 4,086.53$	0.234 ± 0.01	2.48 ± 0.01	-0.466 ± 0.01	$1.38e{-}1 \pm 4.8e{-}3$	$\approx 0\%$	$22.87\% \pm 1.68\%$
	HGEN	$\mathbf{825.6 \pm 22.1}$	$\mathbf{1569.3 \pm 31.3}$	$\mathbf{0.034 \pm 0.003}$	1.634 ± 0.07	$\mathbf{-0.143 \pm 0.08}$	$\mathbf{3.92e{-}2 \pm 7.5e{-}4}$	$0.07\% \pm 0.01\%$	$\mathbf{93.91\% \pm 0.12\%}$
	Real	948	2,114	0.068	1.75	-0.208	N/A	N/A	N/A
IMDB	GraphRNN	1425.47 ± 121.5	142.13 ± 5.87	0.179 ± 0.02	2.97 ± 0.05	0.05 ± 0.04	$1.98e{-}1 \pm 2.61e{-}3$	$9.87\% \pm 0.51\%$	$21.52\% \pm 3.31\%$
	NetGAN	932.5 ± 8.49	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	2.08 ± 0.01	$\mathbf{-0.25 \pm 0.07}$	$1.36e{-}1 \pm 1.89e{-}3$	$7.62\% \pm 0.07\%$	$82.69\% \pm 1.27\%$
	VGAE	635.2 ± 4.16	$7,752.4 \pm 281.32$	0.141 ± 0.01	2.02 ± 0.02	-0.49 ± 0.15	$1.9e{-}1 \pm 2.33e{-}3$	$\approx 0\%$	$42.71\% \pm 1.47\%$
	HGEN	$\mathbf{945.2 \pm 11.54}$	26.0 ± 3.28	$3.56e{-}3 \pm 3.42e{-}4$	$\mathbf{2.16 \pm 0.01}$	-0.19 ± 0.04	$\mathbf{4.36e{-}2 \pm 4.25e{-}4}$	$2.69\% \pm 0.04\%$	$\mathbf{88.71\% \pm 0.39\%}$
	Real	1,074	1	4.43e−4	2.51	-0.235	N/A	N/A	N/A
DBLP	NetGAN	$10,353 \pm 72.71$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	3.308 ± 0.41	-0.059 ± 0.03	$5.03e{-}1 \pm 2.1e{-}2$	$5.48\% \pm 0.32\%$	$\mathbf{72.51\% \pm 0.32\%}$
	VGAE	$3,771 \pm 236.29$	1214.69 ± 452.61	0.271 ± 0.06	1.579 ± 0.07	-0.44 ± 0.11	$8.71e{-}2 \pm 1.77e{-}3$	$\approx 0\%$	$17.26\% \pm 0.41\%$
	HGEN	$\mathbf{5,163 \pm 21.41}$	1068 ± 12.83	0.018 ± 0.001	$\mathbf{1.793 \pm 0.21}$	$\mathbf{-0.157 \pm 0.03}$	$\mathbf{5.82e{-}3 \pm 1.67e{-}4}$	$1.55\% \pm 0.09\%$	$66.59\% \pm 0.17\%$
	Real	5,513	0.0	0.0	1.855	-0.201	N/A	N/A	N/A

TABLE I: Performance evaluation over compared baselines. The *Real* rows include the values of real graphs, while the rest are the evaluation results of different algorithms. The best performance (the closest to real value) achieved under each metric for a particular dataset is highlighted in bold font. Note that we do not include GraphVAE in datasets with (≥ 500) nodes and GraphRNN in datasets with ($\geq 10,000$) nodes because the programs return errors.

(a) Meta-path Length Ratio (b) Meta-path Patterns - Length 1 (c) Meta-path Patterns - Length 2 (d) Meta-path Patterns - Length 3

(e) Meta-path Length Ratio (f) Meta-path Patterns - Length 1 (g) Meta-path Patterns - Length 2 (h) Meta-path Patterns - Length 3
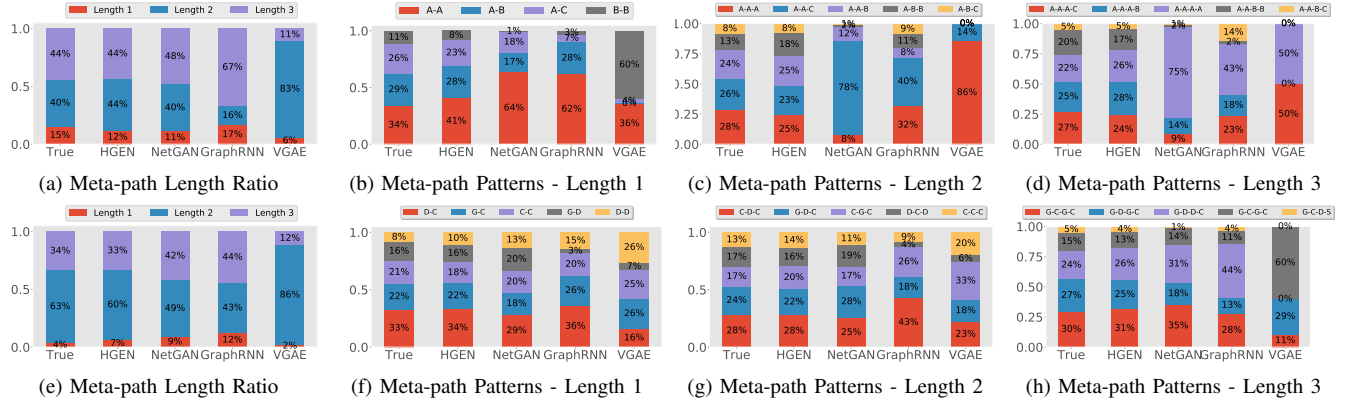
Fig. 5: The meta-path distribution comparison. 5a - 5d and 5e - 5h are the generated meta-path length distribution along with frequent meta-path patterns distribution with length 1, 2, 3 for Syn_500 dataset and PubMed dataset, respectively.

	HGEN-S	HGEN-E	HGEN-A	HGEN	Real
LCC	1563.76	824.14	819.32	**825.6**	948
TC	1453.23	784.34	863.53	**1569.3**	2114
Clustering Coef.	0.026	0.015	0.016	**0.034**	0.068
Power Law Coef.	1.649	**1.652**	1.621	1.634	1.75
Assortativity	-0.09	-0.132	-0.131	$\mathbf{-0.143}$	-0.208
Node Degree Dist.	**0.0354**	0.0388	0.0515	0.0392	N/A

TABLE II: Ablation Study in PubMed Dataset

purely memorize the seen heterogeneous walks in the training data. In contrast, GraphRNN has a higher EO rate, indicating GraphRNN regenerates graphs it saw during training. In

addition, VGAE achieves the lowest EO rate since it fails to generate realistic heterogeneous graphs. For Uniqueness, HGEN also exceeds other one-shot and sequential based algorithms by an evident margin, which demonstrates the diversity of the generated graphs.

Preservation of Graph Semantic Properties To further demonstrate the performance of HGEN, we evaluate the performance of meta-path distribution preservation with other baselines. Specifically, we measure the meta-path distribution from two aspects: 1) the overall meta-path length ratio preservation in generated graphs and 2) frequent meta-path patterns
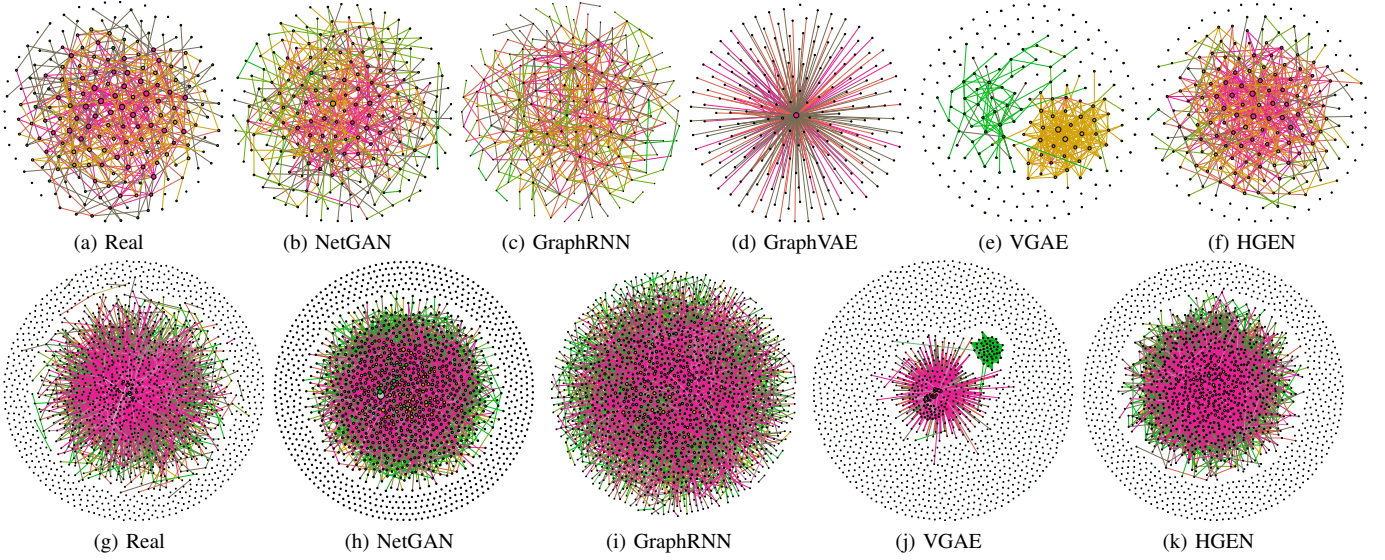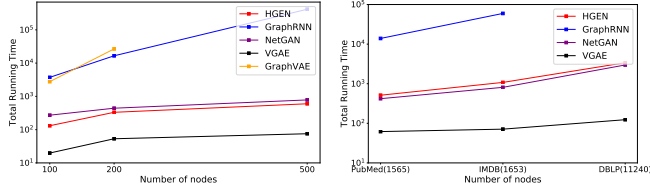
Fig. 6: Figure 6a - 6f are the generated graph of the Syn_200 dataset, and 6g - 6k are the generated graphs of the PubMed dataset. (better to see with color)

(a) Synthetic Dataset Running Time   (b) Real Dataset Running Time

Fig. 7: Running time comparison of different models in both synthetic and real world datasets. It is clear that GraphVAE is not scalable in generating graphs with more than 200 nodes. GraphRNN also fails in generating large graphs (with more than $10,000$ nodes). The proposed HGEN exhibits a linear running time growth in terms of the growth of graph size.

under each length. The results of Syn_500 and PubMed datasets are illustrated in Figure 5 (the rest of the experiments are shown in Appendix). In general, all the methods can approximately maintain the meta-path length ratio except for VGAE. However, HGEN can constantly achieve a better performance as shown in Figure 5a and 5e. 2) As shown in Figure 5b - 5d and 5f - 5h, HGEN can outperform other methods by at least 10% in preserving the ratio of specific meta-path patterns under each length, which is expected since HGEN is able to learn and maintain the meta-path distribution from the observed graphs while others cannot.

D. Ablation Study

We further conduct ablation studies on the PubMed dataset to evaluate the effect of different components in HGEN, and the results are exhibited in Table II. The ablative experiments are conducted based on each of the essential components in our architecture. Specifically, we select a single large heterogeneous walk length - 8 to replace the heterogeneous walk length 1, 2, and 3 in our model, and the resulting model is called HGEN-S. We also independently remove the heterogeneous

node embedding to let the generator uniformly sample the next node, and the resulting model is named HGEN-E. Lastly, we replace the heterogeneous graph assembler with a probability-based graph assembler, namely HGEN-A.

As shown in Table II, all the ablative models achieve similar results in node-level metrics like Powerlaw Coef., Assortativity, which is because HGEN can well capture this node-level information through learning the heterogeneous walk distribution. Other than that, we observe: 1) HGEN-S can construct a larger sub-graph since the length of the heterogeneous walk is largely greater than HGEN, but the large subgraph doe not makes any improvements in terms of capturing the heterogeneous structural information. The reason is there are rarely long meta-path in the heterogeneous graph since longer meta-paths are highly redundant because of the shared sub-parts [10]. We instead choose 1, 2, and 3 as our meta-path lengths to make the whole generation more flexible. 2) removing the heterogeneous node embedding would make HGEN-E hard to capture the local graph structure since HGEN relies on the encoded neighborhood information to make the node sampling be aware of the local structure. 3) as shown in the node degree distribution evaluation, replacing the heterogeneous graph assembler with a probabilistic graph assembler would cause HGEN-A hard to capture the latent heterogeneous node distribution because it uniformly samples edges from the generated walks and completely neglects the generated meta-path information. However, HGEN takes meta-paths as a basic unit to sample edges so that it can effectively preserve the overall distribution of meta-paths as proved in Theorem 1. Therefore, the node degree distribution under each type can be well preserved.

E. Running Time Comparison

The results of our running time experiments are shown in Figure 7. The running times on both synthetic and real-world

datasets including both training and inference time are shown with respect to the growth of number of nodes in both synthetic and real-world datasets. All running times are in $log10$ scale. As shown in both figures, random-walk-based generative models (HGEN and NetGAN) have a constant running time growth in terms of number of nodes, which is especially important when dealing with large graphs. Even though VGAE is much faster regarding running time, it is indeed a representation learning framework based on GCN and lacks of the ability of generating realistic heterogeneous graphs, and the results are also reflected in Table I. Both GraphRNN and GraphVAE fail to compare with HGEN in model scalability because their designs require at least $O(|\mathcal{V}|^2)$ to process the transformed node sequence and adjacency matrix.

F. Graph Visualization

Since it is nearly impossible to judge whether a graph is realistic only by statistics, we visualize the generated graph to further demonstrate the performance of HGEN (Figure 6). Visually, HGEN looks the most similar, while both GraphVAE and VGAE is the most dissimilar. This result is consistent with the quantitative results obtained in Table I. For one-shot based generative models, GraphVAE and VGAE, they fail to capture the structural similarity of the observed heterogeneous graph. For the sequential-based and random walk based graph generative methods, GraphRNN and NetGAN can successfully mimic the structure similarity but fail to preserve the global heterogeneous graph properties (e.g., overall meta-path ratio).

VI. Conclusion

In this paper, we propose a novel framework - HGEN for heterogeneous graph generation, which can jointly capture the semantic, structural, and global distributions of heterogeneous graphs. Our framework consists of a novel heterogeneous walk generator that can hierarchically generate meta-path instances (namely heterogeneous walk) and a heterogeneous graph assembler that can construct new graphs by sampling from the generated heterogeneous walks in a stratified manner. Extensive experiments on synthetic and real-world datasets demonstrate the advantages of HGEN over existing deep generative models in terms of preserving both graph statistical and heterogeneous specified properties.

Acknowledgement

References

[1] M. Sun and P. Li, "Graph to graph: a topology aware approach for graph structures learning and generation," in *Proc. of the AISTATS*, 2019, pp. 2946–2955.

[2] J. You, R. Ying, X. Ren, W. Hamilton, and J. Leskovec, "Graphrnn: Generating realistic graphs with deep auto-regressive models," in *Proc. of the ICML*, 2018.

[3] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph transformer networks," in *Proc. of the NIPS*, 2019, pp. 11 983–11 993.

[4] L. Wu, P. Cui, J. Pei, and L. Zhao, *Graph Neural Networks: Foundations, Frontiers, and Applications*. Singapore: Springer, 2021.

[5] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *Trans. on the NNLS*, 2020.

[6] X. Guo and L. Zhao, "A systematic survey on deep generative models for graph generation," *arXiv preprint arXiv:2007.06686*, 2020.

[7] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *Proc. of the ICDM*.

[8] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao, "Link prediction and recommendation across heterogeneous social networks," in *Proc. of the ICDM*, 2012, pp. 181–190.

[9] J.-Y. Li and M.-Y. Yeh, "On sampling type distribution from heterogeneous social networks," in *Proc. of the PAKDD*, 2011, pp. 111–122.

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

[11] Y. Sun and J. Han, "Meta-path-based search and mining in heterogeneous information networks," *Tsinghua Science and Technology*, vol. 18, no. 4, pp. 329–338, 2013.

[12] 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," *Trans. on the TKDD*, vol. 7, no. 3, pp. 1–23, 2013.

[13] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proc. of the KDD*, 2017, pp. 135–144.

[14] T.-y. Fu, W.-C. Lee, and Z. Lei, "Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proc. of the CIKM*, 2017, pp. 1797–1806.

[15] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *Proc. of the WebConf*, 2019, pp. 2022–2032.

[16] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proc. of the WebConf*, 2020, pp. 2704–2710.

[17] C. Yang, M. Liu, F. He, X. Zhang, J. Peng, and J. Han, "Similarity modeling on heterogeneous networks via automatic path discovery," in *Proc. of the ECML-PKDD*, 2018.

[18] C. Yang, Y. Feng, P. Li, Y. Shi, and J. Han, "Meta-graph based hin spectral embedding: Methods, analyses, and insights," in *Proc. of the ICDM*, 2018.

[19] C. Yang, J. Zhang, and J. Han, "Neural embedding propagation on heterogeneous networks," in *Proc. of the ICDM*, 2019.

[20] C. Yang, Y. Xiao, Y. Zhang, Y. Sun, and J. Han, "Heterogeneous network representation learning: A unified framework with survey and benchmark," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[21] Y. Fan, Y. Ye, Q. Peng, J. Zhang, Y. Zhang, X. Xiao, C. Shi, Q. Xiong, F. Shao, and L. Zhao, "Metagraph aggregated heterogeneous graph neural network for illicit traded product identification in underground market," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 132–141.

[22] W. Zhang, B. Paudel, L. Wang, J. Chen, H. Zhu, W. Zhang, A. Bernstein, and H. Chen, "Iteratively learning embeddings and rules for knowledge graph reasoning," in *Proc. of the WebConf*, 2019, pp. 2366–2377.

[23] C. Shi, Z. Zhang, P. Luo, P. S. Yu, Y. Yue, and B. Wu, "Semantic path based personalized recommendation on weighted heterogeneous information networks," in *Proc. of the CIKM*, 2015, pp. 453–462.

[24] A. Gupta, "Generating large-scale heterogeneous graphs for benchmarking," in *Specifying Big Data Benchmarks*, 2012, pp. 113–128.

[25] X. Guo, L. Zhao, C. Nowzari, S. Rafatirad, H. Homayoun, and S. M. P. Dinakarrao, "Deep multi-attributed graph translation with node-edge co-evolution," in *Proc. of the ICDM*, 2019, pp. 250–259.

[26] A. Bojchevski, O. Shchur, D. Zügner, and S. Günnemann, "Netgan: Generating graphs via random walks," in *Proc. of the ICML*, 2018, pp. 610–619.

[27] M. Simonovsky and N. Komodakis, "Graphvae: Towards generation of small graphs using variational autoencoders," in *Proc. of the ICANN*, 2018, pp. 412–422.

[28] C. Yang, H. Wang, K. Zhang, L. Chen, and L. Sun, "Secure deep graph generation with link differential privacy," in *Proc. of the IJCAI*, 2021.

[29] V. Caridá, A. Jalilifard, A. Mansano, and R. Cristo, "Can netgan be improved on short random walks?" in *Proc. of the BRACIS*, 2019.

[30] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv:1611.07308*, 2016.

[31] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *Trans. on the TKDE*, vol. 29, no. 1, pp. 17–37, 2016.

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

[33] N. De Cao and T. Kipf, "Molgan: An implicit generative model for small molecular graphs," *arXiv preprint arXiv:1805.11973*, 2018.

[34] H. Wang, J. Wang, J. Wang, M. Zhao, W. Zhang, F. Zhang, X. Xie, and M. Guo, "Graphgan: Graph representation learning with generative adversarial nets," in *Proc. of the AAAI*, vol. 32, no. 1, 2018.

[35] C. Yang, P. Zhuang, W. Shi, A. Luu, and P. Li, "Conditional structure generation through graph variational generative adversarial nets," in *Proc. of the NIPS*, 2019, pp. 1340–1351.

[36] S. Honda, H. Akita, K. Ishiguro, T. Nakanishi, and K. Oono, "Graph residual flow for molecular graph generation," *arXiv preprint arXiv:1909.13521*, 2019.

[37] N. Goyal, H. V. Jain, and S. Ranu, "Graphgen: A scalable approach to domain-agnostic labeled graph generation," in *Proc. of the WebConf*, 2020, pp. 1253–1263.

[38] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[39] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. of the ICML*, 2017, pp. 214–223.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. of the NeurIPS*, 2014.