

Link Prediction on Latent Heterogeneous Graphs

Trung-Kien Nguyen*
Singapore Management University
Singapore
tknguyen@smu.edu.sg

Zemin Liu*[†]
National University of Singapore
Singapore
zeminliu@nus.edu.sg

Yuan Fang
Singapore Management University
Singapore
yfang@smu.edu.sg

ABSTRACT

On graph data, the multitude of node or edge types gives rise to heterogeneous information networks (HINs). To preserve the heterogeneous semantics on HINs, the rich node/edge types become a cornerstone of HIN representation learning. However, in real-world scenarios, type information is often noisy, missing or inaccessible. Assuming no type information is given, we define a so-called *latent heterogeneous graph* (LHG), which carries latent heterogeneous semantics as the node/edge types cannot be observed. In this paper, we study the challenging and unexplored problem of link prediction on an LHG. As existing approaches depend heavily on type-based information, they are suboptimal or even inapplicable on LHGs. To address the absence of type information, we propose a model named LHGNN, based on the novel idea of semantic embedding at node and path levels, to capture latent semantics on and between nodes. We further design a personalization function to modulate the heterogeneous contexts conditioned on their latent semantics w.r.t. the target node, to enable finer-grained aggregation. Finally, we conduct extensive experiments on four benchmark datasets, and demonstrate the superior performance of LHGNN.

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations**; • **Information systems** → *Data mining*.

KEYWORDS

Latent heterogeneous graph, link prediction, graph neural networks

ACM Reference Format:

Trung-Kien Nguyen*, Zemin Liu*[†], and Yuan Fang. 2023. Link Prediction on Latent Heterogeneous Graphs. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*, May 1–5, 2023, Austin, TX, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3543507.3583284>

*Co-first authors with equal contribution.

[†]Corresponding author. Work partly done at Singapore Management University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WWW '23, May 1–5, 2023, Austin, TX, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9416-1/23/04...\$15.00
<https://doi.org/10.1145/3543507.3583284>

1 INTRODUCTION

Objects often interact with each other to form graphs, such as the Web and social networks. The prevalence of graph data has catalyzed graph analysis in various disciplines. In particular, link prediction [43] is a fundamental graph analysis task, enabling widespread applications such as friend suggestion in social networks, recommendation in e-commerce graphs, and citation prediction in academic networks. In these real-world applications, the graphs are typically *heterogeneous* as opposed to *homogeneous*, also known as Heterogeneous Information Networks (HINs) [39], in which multiple types of nodes and edges exist. For instance, the academic HIN shown in the top half of Fig. 1(a) interconnects nodes of three types, namely, Author (A), Paper (P) and Conference (C), through different types of edges such as “writes/written by” between author and paper nodes and “publishes/published in” between conference and paper nodes, *etc.* The multitude of node and edge types in HINs implies rich and diverse semantics on and between nodes, which opens up great opportunities for link prediction.

A crucial step for link prediction is to derive features from an input graph. Recent literature focuses on graph representation learning [4, 37], which aims to map the nodes on the graph into a low-dimensional space that preserves the graph structures. Various approaches exist, ranging from the earlier shallow embedding models [10, 27, 32] to more recent message-passing graph neural networks (GNNs) [11, 18, 34, 38]. Representation learning on HINs generally follows the same paradigm, but aims to preserve the heterogeneous semantics in addition to the graph structures in the low-dimensional space. To express heterogeneous semantics, existing work resorts to type-based information, including simple node/edge types (*e.g.*, an author node carries different semantics from a paper node), and type structures like metapath [30] (*e.g.*, the metapath A-P-A implies two authors are collaborators, whereas A-P-C-P-A implies two authors in the same field; see Sect. 3 for the metapath definition). Among the state-of-the-art heterogeneous GNNs, while hinging on the common operation of message passing, some exploit node/edge types [12, 16, 22, 41] and others employ type structures [9, 28, 35].

Our problem. The multitude of node or edge types gives rise to rich heterogeneous semantics on HINs, and forms the key thesis of HIN representation learning [39]. However, in many real-world scenarios, type information is often noisy, missing or inaccessible. One reason is type information does not exist explicitly and has to be deduced. For instance, when extracting entities and their relations from texts to construct a knowledge graph, NLP techniques are widely used to classify the extractions into different types, which can be noisy. Another reason is privacy and security, such that the nodes in a network may partially or fully hide their identities and types. Lastly, even on an apparently homogeneous

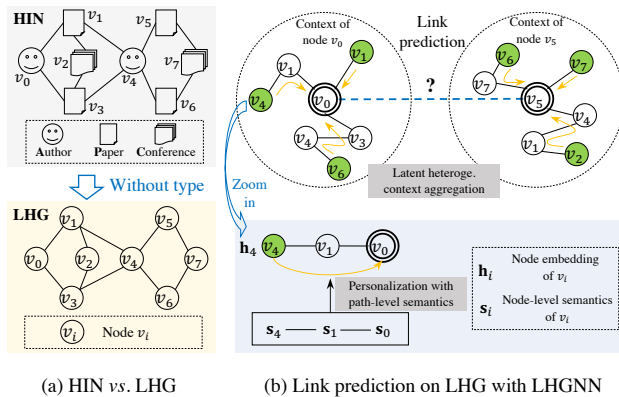


Figure 1: Illustration of our problem and approach. (a) Comparison of HIN and LHG. (b) Key insights of our approach.

graph, such as a social network which only consists of users and their mutual friendships, could have fine-grained latent types, such as different types of users (e.g., students and professionals) and different types of friendships (e.g., friends, family and colleagues), but we cannot observe such latent types.

Formally, we call those HINs without explicit type information as *Latent Heterogeneous Graphs* (LHGs), as shown in the bottom half of Fig. 1(a). The key difference on an LHG is that, while different types still exist, the type information is completely inaccessible and cannot be observed by data consumers. It implies that LHGs still carry rich heterogeneous semantics that are crucial to effective representation learning, but the heterogeneity becomes latent and presents a much more challenging scenario given that types or type structures can no longer be used. In this paper, we investigate this unexplored problem of *link prediction on LHGs*, which calls for modeling the latent semantics on LHGs as links are formed out of relational semantics between nodes.

Challenges and insights. We propose a novel model for link prediction on LHGs, named Latent Heterogeneous Graph Neural Network (LHGNN). Our general idea is to develop a latent heterogeneous message-passing mechanism on an LHG, in order to exploit the latent semantics between nodes for link prediction. More specifically, we must address two major challenges.

First, *how to capture the latent semantics on and between nodes without any type information?* In the absence of explicit type information, we resort to *semantic embedding* at both node and path levels to depict the latent semantics. At the node level, we complement the traditional representation of each node (e.g., h_4 in Fig. 1(b), which we call the *primary embedding*) with an additional semantic embedding (e.g., s_4). While the primary embedding can be regarded as a blend of content and structure information, the semantic embedding aims to distill the more subtle semantic information (i.e., latent node type and relation types a node tends to associate with), which could have been directly expressed if explicit types were available. Subsequently, at the path level, we can learn a semantic path embedding based on the node-level semantic embeddings in a path on an LHG, e.g., s_4, s_1 and s_0 in the path $v_4 - v_1 - v_0$ in Fig. 1(b). The path-level semantic embeddings aim to capture

the latent high-order relational semantics between nodes, such as the collaborator relation between authors v_4 and v_0 in Fig. 1(a), to mimic the role played by metapaths such as A-P-A.

Second, as context nodes of a target node often carry latent heterogeneous semantics, *how to differentiate them for finer-grained context aggregation?* We propose a learnable *personalization* function to modulate the original message from each context node. The personalization hinges on the semantic path embedding between each context node and the target node as the key differentiator of heterogeneous semantics carried by different context nodes. For illustration, refer to the top half of Fig. 1(b), where the green nodes (e.g., v_1, v_4 and v_6) are the context nodes of the doubly-circled target node (e.g., v_0). These context nodes carry latent heterogeneous semantics (e.g., v_1 is a paper written by v_0 , v_4 is a collaborator of v_0 , and v_6 is a related paper v_0 might be interested in), and thus can be personalized by the semantic path embedding between each context and the target node before aggregating them.

Contributions. In summary, our contributions are three-fold. (1) We investigate a novel problem of link prediction on latent heterogeneous graphs, which differs from traditional HINs due to the absence of type information. (2) We propose a novel model LHGNN based on the key idea of semantic embedding to bridge the gap for representation learning on LHGs. LHGNN is capable of inferring both node- and path-level semantics, in order to personalize the latent heterogeneous contexts for finer-grained message passing within a GNN architecture. (3) Extensive experiments on four real-world datasets demonstrate the superior performance of LHGNN in comparison to the state-of-the-art baselines.

2 RELATED WORK

Graph neural networks. GNNs have recently become the mainstream for graph representation learning. Modern GNNs typically follow a message-passing scheme, which derives low-dimensional embedding of a target node by aggregating messages from context nodes. Different schemes of context aggregation have been proposed, ranging from simple mean pooling [11, 18] to neural attention [34] and other neural networks [38].

For representation learning on HINs, a plethora of heterogeneous GNNs have been proposed. They depend on type-based information to capture the heterogeneity, just as earlier HIN embedding approaches [5, 8, 14]. On one hand, many of them leverage simple node/edge types. HetGNN [41] groups random walks by node types, and then applies bi-LSTM to aggregate node- and type-level messages. HGT [16] employs node- and edge-type dependent parameters to compute the heterogeneous attention over each edge. Simple-HGN [22] extends the edge attention with a learnable edge type embedding, whereas HetSANN [12] employs a type-aware attention layer. On the other hand, high-order type structures such as meta-path [30] or meta-graph [6] have also been used. HAN [35] uses meta-paths to build homogeneous neighbor graphs to facilitate node- and semantic-level attentions in message aggregation, whereas MAGNN [9] proposes several meta-path instance encoders to account for intermediate nodes in a meta-path instance. Meta-GNN [28] differentiates context nodes based on meta-graphs. Another work Space4HGNN [45] proposes a unified design space to build heterogeneous GNNs in a modularized manner,

which can potentially utilize various type-based information. Despite their effectiveness on HINs, they cannot be applied to LHGs due to the need of explicit type information.

Knowledge graph embedding. A knowledge graph consists of a large number of relations between head and tail entities. Translation models [3, 19, 36] are popular approaches that treat each relation as a translation in some vector space. For example, TransE [3] models a relation as a translation between head and tail entities in the same embedding space, while TransR [19] further maps entities into multiple relation spaces to enhance semantic expressiveness. Separately, RotatE [31] models each relation as a rotation from head to tail entities in complex space, and is able to capture different relation patterns such as symmetry and inversion. These models require the edge type (*i.e.*, relation) as input, which is not available on an LHG. Compared to heterogeneous GNNs, they do not utilize entity features or model multi-hop interactions between entities, which can lead to inferior performance on HINs.

Predicting missing types. A few studies [13, 25] aim to predict missing entity types on knowledge graphs. However, a recent work [42] shows that these approaches tend to propagate type prediction errors on the graph, which harms the performance of other tasks like link prediction. Therefore, RPGNN [42] proposes a relation encoder, which is adaptive for each node pair to handle the missing types. However, all of them still require partial type information from a subset of nodes and edges for supervised training, which makes them infeasible in the LHG setting.

3 PRELIMINARIES

We first review or define the core concepts in this work.

Heterogeneous Information Networks (HIN). An HIN [29] is defined as a graph $G = (V, E, T, R, \psi, \phi)$, where V denotes the set of nodes and T denotes the set of node types, E denotes the set of edges and R denotes the set of edge types. Moreover, $\psi : V \rightarrow T$ and $\phi : E \rightarrow R$ are functions that map each node and edge to their types in T and R , respectively. G is an HIN if $|T| + |R| > 2$.

Latent Heterogeneous Graph (LHG). An LHG is an HIN $G = (V, E, T, R, \psi, \phi)$ such that the types T, R and mapping functions ψ, ϕ are *not accessible*. That is, we only observe a *homogeneous* graph $G' = (V, E)$ without knowing T, R, ψ, ϕ .

Metapath and latent metapath. On an HIN, a metapath P is a sequence of node and edge types [30]: $P = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} T_{l+1}$, such that $T_i \in T$ and $R_i \in R$. As an edge type $R_i \in R$ represents a relation, a metapath represents a composition of relations $R_1 \circ R_2 \circ \dots \circ R_l$. Hence, metapaths can capture complex, high-order semantics between nodes. A path $p = (v_1, v_2, \dots, v_{l+1})$ on the HIN is an instance of metapath P if $\psi(v_i) = T_i$ and $\phi(\langle v_i, v_{i+1} \rangle) = R_i$. As shown in the top half of Fig. 1(a), an example metapath is Author-Paper-Author (A-P-A), implying the collaborator relation between authors. Instances of this metapath include $v_0-v_1-v_4$ and $v_0-v_3-v_4$, signifying that v_0 and v_4 are collaborators.

On an LHG the metapaths become latent too, as the types T_i and R_i are not observable. Generally, any path $p = (v_1, v_2, \dots, v_{l+1})$ on an LHG is an instance of some latent metapath, which carries latent semantics representing an unknown composition of relations between the starting node v_1 and end node v_{l+1} .

Link prediction on LHG. Given a query node, we rank other nodes by their probability of forming a link with the query. The difference lies in the input graph, where we are given an LHG.

4 PROPOSED METHOD: LHGNN

In this section, we introduce the proposed method LHGNN for link prediction on latent heterogeneous graphs.

4.1 Overall Framework

We start with the overall framework of LHGNN, as presented in Fig. 2. An LHG is given as input as illustrated in Fig. 2(a), which is fed into an LHGNN layer in Fig. 2(b). Multiple layers can be stacked, and the last layer would output the node representations, to be further fed into a link encoder for the task of link prediction as shown in Fig. 2(c). More specifically, the LHGNN layer is our core component, which consists of two sub-modules: a *semantic embedding* sub-module to learn node-level and path-level latent semantics, and a *latent heterogeneous context aggregation* sub-module to aggregate messages for the target node. We describe each sub-module and the link prediction task in the following.

4.2 Semantic Embedding

Semantic embedding aims to model both node- and path-level latent semantics, as illustrated in Fig. 2(b1).

Node-level semantic embedding. For each node v , alongside its primary embedding \mathbf{h}_v , we propose an additional semantic embedding \mathbf{s}_v . Similar to node embeddings on homogeneous graphs, the primary embeddings intend to capture the overall content and structure information of nodes. However, on an LHG, the content of a node contains not only concrete topics and preferences, but also subtle semantic information inherent to nodes of each latent type (*e.g.*, node type, and potential single or multi-hop relations a node tends to be part of). Hence, we propose a semantic encoder to locate and distill the subtle semantic information from the primary embeddings to generate semantic embeddings, which will be later employed for link prediction. Note that this is different from disentangled representation learning [1], which can be regarded as disentangling a mixture of latent topics.

Specifically, in the l -th layer, given primary embeddings from the previous layer¹ $(\mathbf{h}_{v_1}^{l-1}, \mathbf{h}_{v_2}^{l-1}, \dots)$, a semantic encoder generates the corresponding semantic embeddings $(\mathbf{s}_{v_1}^l, \mathbf{s}_{v_2}^l, \dots)$. For each node v , the semantic encoder f_s extracts the semantic embedding \mathbf{s}_v^l from its primary embedding \mathbf{h}_v^{l-1} : $\mathbf{s}_v^l = f_s(\mathbf{h}_v^{l-1}; \theta_s^l) \in \mathbb{R}^{d_s^l}$. While the function $f_s(\cdot; \theta_s^l)$ can take many forms, we simply materialize it as a fully connected layer:

$$\mathbf{s}_v^l = \text{LEAKYRELU}(\mathbf{W}_s^l \mathbf{h}_v^{l-1} + \mathbf{b}_s^l), \quad (1)$$

where $\mathbf{W}_s^l \in \mathbb{R}^{d_s^l \times d_h^{l-1}}$ and $\mathbf{b}_s^l \in \mathbb{R}^{d_s^l}$ are the parameters of the encoder, *i.e.*, $\theta_s^l = \{\mathbf{W}_s^l, \mathbf{b}_s^l\}$. Since the semantic embedding only distill the subtle semantic information from the primary embedding, it needs much fewer dimensions, *i.e.*, $d_s^l \ll d_h^l$.

Path-level semantic embedding. A target node is often connected with many context nodes through paths. On an LHG, these

¹When $l = 0$, the primary embedding \mathbf{h}_v^0 is set to the input node features of v_i .

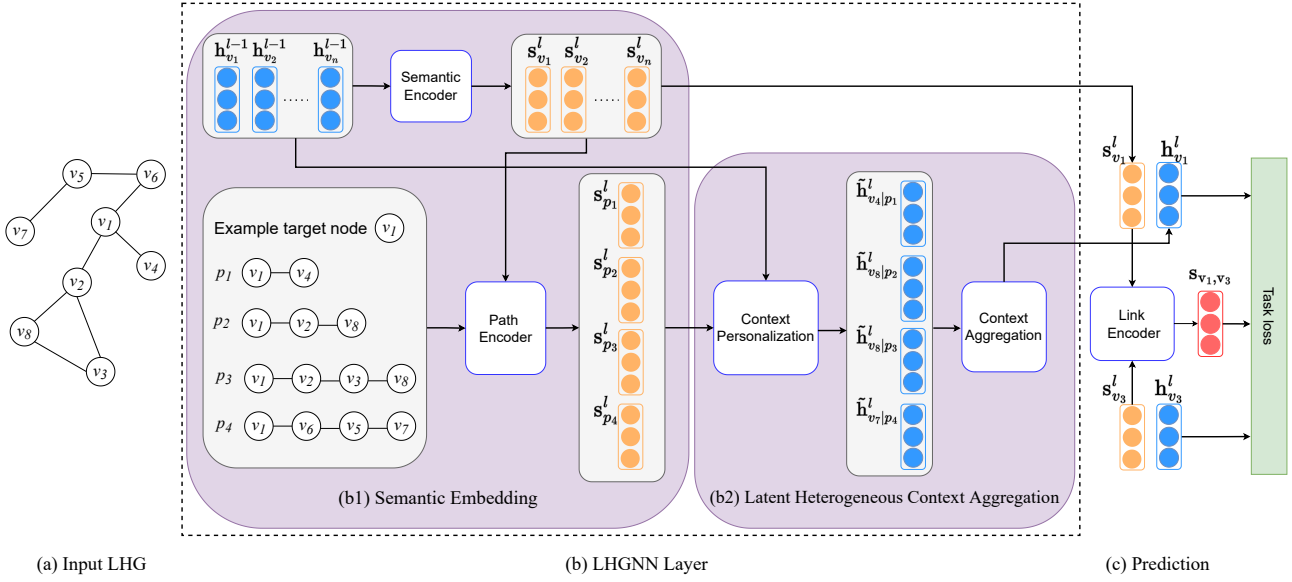


Figure 2: Overall framework of LHGNN.

paths may carry different latent semantics by virtue of the heterogeneous multi-hop relations between nodes. In an HIN, to capture the heterogeneous semantics from different context nodes, metapaths have been a popular tool. For example, in the top half of Fig. 1(a), the HIN consists of three types of nodes. There exist different metapaths that capture different semantics between authors: A-P-A for authors who have directly collaborated, or A-P-C-P-A for authors who are likely in the same field, *etc.* However, on an LHG, we do not have access to the node types, and thus are unable to define or use any metapath. Thus, we employ a path encoder to fuse the node-level semantic embeddings associated with a path into a path-level embedding. The path-level semantic embeddings attempt to mimic the role of metapaths on an HIN, to capture the latent heterogeneous semantics between nodes.

Concretely, we first perform random walks to sample a set of paths. Starting from each target node, we sample N random walks of length L_{\max} , e.g., $\{p_1, \dots, p_4\}$ for v_1 as shown in Fig. 2(b). Then, for each sampled walk, we truncate it into a shorter path with length $L \leq L_{\max}$. These paths of varying lengths can capture latent semantics of different ranges, and serve as the instances of different latent metagraphs.

Next, for each path p_i , a path encoder encodes it to generate a semantic path embedding $s_{p_i}^l$. Let P_v denote the set of sampled paths starting from a target node v . If there exists a path $p_i \in P_v$ such that p_i ends at node u , we call u a *context node* of v or simply *context* of v . For instance, given the set of paths $P_{v_1} = \{p_1, \dots, p_4\}$ for the target node v_1 in Fig. 2(b1), the contexts are v_4, v_8, v_7 , which carry different semantics for v_1 . In the l -th layer, we employ a path encoder f_p to embed each path p_i into a semantic path embedding $s_{p_i}^l \in \mathbb{R}^{d_s}$ based on the node-level semantic embeddings of each node v_j in the path.

$$s_{p_i}^l = f_p(\{s_{v_j}^l \mid v_j \text{ in the path } p_i\}), \quad (2)$$

where f_p can take many forms, ranging from simple pooling to recurrent neural networks [24] and transformers [33]. As the design of path encoder is not the focus of this paper, we simply implement it as a mean pooling, which computes the mean of the node-level semantic embeddings as the path-level embedding.

REMARK. It is possible or even likely to have multiple paths between the target and context nodes. Intrinsicly, these paths are instances of one or more latent metapaths, which bear a two-fold advantage. First, as each latent metapath depicts a particular semantic relationship between two nodes, having plural latent metapaths could capture different semantics between the two nodes. This is more general than many heterogeneous GNNs [9, 35] on a HIN, which rely on a few handcrafted metapaths. Second, it is much more efficient to sample and process paths than subgraphs [17]. Although a metagraph can express more complex semantics than a metapath [44], the combination of multiple metapaths is a good approximation especially given the efficiency trade-off.

4.3 Latent Heterogeneous Context Aggregation

To derive the primary embedding of the target node, the next step is to perform latent heterogeneous context aggregation for the target node. The aggregation generally follows a message-passing GNN architecture, where messages from context nodes are passed to the target node. For example, consider the target node v_1 shown in Fig. 2(b2). The contexts of v_1 include $\{v_4, v_8, v_7\}$ based on the paths p_1, \dots, p_4 , and their messages (i.e., embeddings) would be aggregated to generate the primary embedding of v_1 .

However, on an LHG, the contexts of a target node carry latent heterogeneous semantics. Thus, these heterogeneous contexts shall be differentiated before aggregating them, in order to preserve the latent semantics in fine granules. Note that on an HIN, the given node/edge types or type structures can be employed as explicit differentiators for the contexts. In contrast, the lack of type

information on an LHG prevents the adoption of the explicit differentiators, and we resort to path semantic embeddings as the basis to differentiate the messages from heterogeneous contexts. That is, in each layer, we *personalize* the message from each context node conditioned on the semantic path embeddings between the target and context node. The personalized context messages are finally aggregated to generate the primary embedding of the target node, *i.e.*, the output of the current layer.

Context personalization. Consider a target node v , a context node u , and their connecting path p . In particular, u can be differentiated from other contexts of v given their connecting path p , which is associated with a unique semantic path embedding. Note that there may be more than one paths connecting v and u . In that case, we treat u as multiple contexts, one instance for each path, as each path carries different latent semantics and the corresponding context instance needs to be differentiated.

Specifically, in the l -th layer, we personalize the message from u to v with a learnable transformation function τ , which modulates u 's original message \mathbf{h}_u^{l-1} (*i.e.*, its primary embedding from the previous layer) into a personalized message $\tilde{\mathbf{h}}_{u|p}^l$ conditioned on the path p between u and v . That is,

$$\tilde{\mathbf{h}}_{u|p}^l = \tau(\mathbf{h}_u^{l-1}, \mathbf{s}_p^l; \theta_\tau^l), \quad (3)$$

where the transformation function $\tau(\cdot; \theta_\tau^l)$ is learnable with parameters θ_τ^l . We implement τ using a layer of feature-wise linear modulation (FiLM) [20, 26], which enables the personalization of a message (*e.g.*, \mathbf{h}_u^{l-1}) conditioned on arbitrary input (*e.g.*, \mathbf{s}_p^l). The FiLM layer learns to perform scaling and shifting operations to modulate the original message:

$$\tilde{\mathbf{h}}_{u|p}^l = (\gamma_p^l + \mathbf{1}) \odot \mathbf{h}_u^{l-1} + \beta_p^l, \quad (4)$$

where $\gamma_p^l \in \mathbb{R}^{d_h^{l-1}}$ is a scaling vector and $\beta_p^l \in \mathbb{R}^{d_h^{l-1}}$ is a shifting vector, both of which are learnable and specific to the path p . Note that $\mathbf{1}$ is a vector of ones to center the scaling around one, and \odot denotes the element-wise multiplication. To make γ_p^l and β_p^l learnable, we materialize them using a fully connected layer, which takes in the semantic path embedding \mathbf{s}_p^l as input to become conditioned on the path p , as follows.

$$\gamma_p^l = \text{LEAKYRELU}(\mathbf{W}_\gamma^l \mathbf{s}_p^l + \mathbf{b}_\gamma^l), \quad (5)$$

$$\beta_p^l = \text{LEAKYRELU}(\mathbf{W}_\beta^l \mathbf{s}_p^l + \mathbf{b}_\beta^l), \quad (6)$$

where $\mathbf{W}_*^l \in \mathbb{R}^{d_h^{l-1} \times d_s^l}$ and $\mathbf{b}_*^l \in \mathbb{R}^{d_h^{l-1}}$ are learnable parameters, and $\text{LEAKYRELU}(\cdot)$ is the activation function. Note that the parameters of the transformation function τ in layer l boil down to parameters of the fully connected layers, *i.e.*, $\theta_\tau^l = \{\mathbf{W}_\gamma^l, \mathbf{W}_\beta^l, \mathbf{b}_\gamma^l, \mathbf{b}_\beta^l\}$.

Context aggregation. Next, we aggregate the personalized messages from latent heterogeneous contexts into $\mathbf{c}_v^l \in \mathbb{R}^{d_h^{l-1}}$, the aggregated context embedding for the target node v :

$$\mathbf{c}_v^l = \text{MEAN}(\{e^{-\lambda L(p)} \tilde{\mathbf{h}}_{u|p}^l \mid p \in P_v\}), \quad (7)$$

where $L(p)$ gives the length of the path p so that $e^{-\lambda L(p)}$ acts as a weighting scheme to bias toward shorter paths, and $\lambda > 0$ is a hyperparameter controlling the decay rate. We use mean-pooling

as the aggregation function, although other choices such as sum- or max-pooling could also be used.

Note that the self-information of the target node is also aggregated, by defining a self-loop on the target node as a special path. More specifically, given a target node v and its self-loop p , we define $L(p) = 0$ and $\tilde{\mathbf{h}}_{v|p}^l = \mathbf{h}_v^l$, which means the original message of the target node will be included into \mathbf{c}_v^l with a weight of 1.

Finally, based on the aggregated context embedding, we obtain the primary embedding of node v in the l -th layer:

$$\mathbf{h}_v^l = \text{LEAKYRELU}(\mathbf{W}_h^l \mathbf{c}_v^l + \mathbf{b}_h^l), \quad (8)$$

where $\mathbf{W}_h^l \in \mathbb{R}^{d_h^l \times d_h^{l-1}}$ and $\mathbf{b}_h^l \in \mathbb{R}^{d_h^l}$ are learnable parameters.

4.4 Link Prediction

In the following, we discuss the treatment of the link prediction task on an LHG, as illustrated in Fig. 2(c). In particular, we will present a link encoder and the loss function.

Link encoder. For link prediction between two nodes, we design a *link encoder* to capture the potential latent relationships between the two candidates. Given two candidate nodes a and b and their respective semantic embeddings $\mathbf{s}_a, \mathbf{s}_b$ obtained from the last LHGNN layer, the link encoder is materialized in the form of a recurrent unit to generate a pairwise semantic embedding:

$$\mathbf{s}_{a,b} = \tanh(\mathbf{W}\mathbf{s}_b + \mathbf{U}\mathbf{s}_a + \mathbf{b}), \quad (9)$$

where $\mathbf{s}_{a,b} \in \mathbb{R}^{d_h}$ can be interpreted as an embedding of the latent relationships between the two nodes a and b , $\mathbf{W}, \mathbf{U} \in \mathbb{R}^{d_s \times d_h}$, $\mathbf{b} \in \mathbb{R}^{d_h}$ are learnable parameters. Here d_h and d_s are the number of dimensions of the primary and semantic embeddings from the last layer, respectively. Note that $\mathbf{s}_{a,b}$ has the same dimension as the node representations, which can be used as a translation to relate nodes a and b in the loss function in the next part.

Loss function. We adopt a triplet loss for link prediction. For an edge $(a, b) \in E$, we construct a triplet (a, b, c) , where c is a negative node randomly sampled from the graph. Inspired by translation models in knowledge graph [3, 19], b can be obtained by a translation on a and the translation approximates the latent relationships between a and b , *i.e.*, $\mathbf{h}_b \approx \mathbf{h}_a + \mathbf{s}_{a,b}$. Note that \mathbf{h}_v denotes the primary node embedding from the final LHGNN layer. In contrast, since c is a random node unrelated to a , \mathbf{h}_c cannot be approximated by the translation. Thus, given a set of training triplets T , we formulate the following triplet margin loss for the task:

$$\mathcal{L}_{\text{task}} = \frac{1}{|T|} \sum_{(a,b,c) \in T} \max(d(a, b) - d(a, c) + \alpha, 0), \quad (10)$$

where $d(x, y) = \|\mathbf{h}_x + \mathbf{s}_{x,y} - \mathbf{h}_y\|_2$ is the Euclidean norm of the translational errors, and $\alpha > 0$ is the margin hyperparameter.

Besides the task loss, we also add constraints to scaling and shifting in the FiLM layer. During training, the scaling and shifting may become arbitrarily large to overfit the data. To prevent this issue, we restrict the search space by the following loss term on the scaling and shifting vectors.

$$\mathcal{L}_{\text{FiLM}} = \sum_{l=1}^{\ell} \sum_{p \in P} (\|\gamma_p^l\|_2 + \|\beta_p^l\|_2), \quad (11)$$

Table 1: Summary of Datasets.

Attributes	FB15k-237	WN18RR	DBLP	OGB-MAG
# Nodes	14,541	40,943	18,405	100,002
# Edges	310,116	93,003	67,946	1,862,256
# Features	-	-	334	128
# Node types	-	-	3	4
# Edge types	237	11	4	4
Avg(degree)	29.09	3.50	3.55	17.88
# Training	272,115	86,835	54,356	1,489,804
# Validation	17,535	3,034	6,794	186,225
# Testing	20,466	3,134	6796	186,227

where ℓ is the total number of layers and P is the set of all sampled paths. The overall loss is then

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \mu \mathcal{L}_{\text{FiLM}}, \quad (12)$$

where $\mu > 0$ is a hyperparameter to balance the loss terms.

We further present the training algorithm for LHGNN in Appendix A, and give a complexity analysis therein.

5 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the effectiveness of LHGNN on four benchmark datasets.

5.1 Experimental Setup

Datasets. We employ four graph datasets summarized in Table 1. Note that, while all these graphs include node or edge types, we hide such type information to transform them into LHGs.

- *FB15k-237* [31] is a refined subset of Freebase [2], a knowledge graph about general facts. It is curated with the most frequently used relations, in which each node is an entity, and each edge represents a relation.
- *WN18RR* [31] is a refined subset of WordNet [23], a knowledge graph demonstrating lexical relations of vocabulary.
- *DBLP* [40] is an academic bibliographic network, which includes three types of nodes, *i.e.*, paper (P), author (A) and conference (C), as well as four types of edges, *i.e.*, P-A, A-P, P-C, C-P. The node features are 334-dimensional vectors which represent the bag-of-word vectors for the keywords.
- *OGB-MAG* [15] is a large-scale academic graph. It contains four types of nodes, *i.e.*, paper (P), author (A), institution (I) and field (F), as well as four types of edges, *i.e.*, A-I, A-P, P-P, P-F. Features of each paper is a 128-dimensional vector generated by word2vec, while the node feature of the other types is generated by metapath2vec [5] with the same dimension following previous work [7]. In our experiments, we randomly sample a subgraph with around 100K entities from the original graph using breadth first search.

Baselines. For a comprehensive comparison, we employ baselines from three major categories.

- *Classic GNNs*: GCN [18], GAT [34] and GraphSAGE [11], which are classic GNN models for homogeneous graphs.

- *Heterogeneous GNNs*: HAN [35], HGT [16] and Simple-HGN (HGN for short) [22], which are state-of-the-art heterogeneous graph neural networks (HGNNs) taking in an HIN as input.
- *Translation models*: TransE [3] and TransR [19], which are well-known methods for knowledge graph embedding.

Note that the heterogeneous GNNs and translation models require node/edge types as input, to apply them to an LHG, we adopt two strategies: either treating all nodes or edges as one type, or generating *pseudo types*, which we will elaborate later. See Appendix B for more detailed descriptions of the baselines.

Model settings. See Appendix D for the hyperparameters and other settings of the baselines and our method.

5.2 Evaluation of Link Prediction

In this part, we evaluate the performance of LHGNN on the main task of link prediction on LHGs.

Settings. For knowledge graph datasets (FB15k-237 and WN18RR), we use the same training/validation/testing proportion in previous work [31], as given in Table 1. For the other datasets, we adopt a 80%/10%/10% random splitting of the links. Note that the training graphs are reconstructed from only the training links.

We adopt ranking-based evaluation metrics for link prediction, namely, NDCG and MAP [21]. In the validation and testing set, given a ground-truth link (a, b) , we randomly sample another 9 nodes which are not linked to a as negative nodes, and form a candidate list together with node b . For evaluation, we rank the 10 nodes based on their scores w.r.t. node a . For our LHGNN, the score for a candidate link (x, y) is computed as $-\|\mathbf{h}_x + \mathbf{s}_{x,y} - \mathbf{h}_y\|_2$. For classic and heterogeneous GNN models, we implement the same triplet margin loss for them, as given by Eq. (10). The only difference is that $d(x, y)$ is defined by $\|\mathbf{h}_x - \mathbf{h}_y\|_2$ in the absence of semantic embeddings. Similarly, their link scoring function is defined as $-\|\mathbf{h}_x - \mathbf{h}_y\|_2$ for a candidate link (x, y) . Translation models also use the same loss and scoring function as ours, except for replacing our link encoding $\mathbf{s}_{x,y}$ with their type-based relation embedding.

Scenarios of comparison. Since the type information is inaccessible on LHGs, for heterogeneous GNNs and the translation methods, we consider the following scenarios.

The first scenario is to treat all nodes/edges as only one type in the absence of type information.

In the second scenario, we generate *pseudo types*. For nodes, we resort to the K -means algorithm to cluster nodes into K clusters based on their features, and treat the cluster ID of each node as its type. Since the number of clusters or types K is unknown, we experiment with different values. For each heterogeneous GNN or translation model, we use “X- K ” to denote a variant of model X with K pseudo node types, where X is the model name. For instance, HAN-3 means HAN with three pseudo node types. Note that there is no node feature in FB15k-237 and WN18RR. To perform clustering, we use node embeddings by running X-1 first. On the other hand, edge types are derived using the Cartesian product of the node types, resulting in $K \times K$ pseudo edge types. Finally, for HAN which requires metapath, we further construct pseudo metapaths based on the pseudo node types. For each pseudo node type, we employ all metapaths with length two starting and ending

Table 2: Evaluation of link prediction on LHGs. Best is bolded and runner-up underlined; OOM means out-of-memory error.

Methods	FB15k-237		WN18RR		DBLP		OGB-MAG	
	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG
GCN	0.790 ± 0.001	0.842 ± 0.001	0.729 ± 0.002	0.794 ± 0.001	0.879 ± 0.001	0.910 ± 0.001	0.848 ± 0.001	0.886 ± 0.001
GAT	0.786 ± 0.002	0.839 ± 0.001	0.761 ± 0.001	0.818 ± 0.001	<u>0.913</u> ± 0.001	<u>0.936</u> ± 0.001	0.830 ± 0.004	0.872 ± 0.003
GraphSAGE	<u>0.800</u> ± 0.001	<u>0.850</u> ± 0.001	0.728 ± 0.003	0.793 ± 0.002	0.891 ± 0.001	0.918 ± 0.001	<u>0.849</u> ± 0.001	<u>0.887</u> ± 0.001
TransE	0.675 ± 0.001	0.752 ± 0.001	0.511 ± 0.002	0.624 ± 0.001	0.488 ± 0.001	0.605 ± 0.001	0.552 ± 0.001	0.656 ± 0.001
TransR	0.734 ± 0.004	0.798 ± 0.003	0.510 ± 0.002	0.623 ± 0.001	0.565 ± 0.007	0.668 ± 0.005	0.546 ± 0.001	0.652 ± 0.001
HAN	0.725 ± 0.002	0.793 ± 0.002	0.749 ± 0.003	0.810 ± 0.003	0.763 ± 0.005	0.801 ± 0.004	OOM	OOM
HGT	0.782 ± 0.001	0.837 ± 0.001	0.724 ± 0.003	0.791 ± 0.002	0.897 ± 0.001	0.923 ± 0.001	0.835 ± 0.003	0.876 ± 0.002
HGN	0.742 ± 0.002	0.806 ± 0.001	<u>0.802</u> ± 0.002	<u>0.849</u> ± 0.002	0.907 ± 0.003	0.930 ± 0.002	0.818 ± 0.001	0.863 ± 0.001
LHGNN	0.858 ± 0.001	0.893 ± 0.001	0.838 ± 0.003	0.877 ± 0.002	0.932 ± 0.003	0.949 ± 0.002	0.879 ± 0.001	0.909 ± 0.001

Table 3: Evaluation of link prediction on LHGs with pseudo types for heterogeneous GNNs and translation models.

Methods	FB15k-237		WN18RR		DBLP		OGB-MAG	
	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG
TransE-3	0.693	0.767	0.510	0.623	0.599	0.693	0.568	0.670
TransE-10	0.701	0.773	0.519	0.630	0.677	0.754	0.599	0.694
TransR-3	0.749	0.810	0.485	0.604	0.585	0.683	0.599	0.695
TransR-10	0.727	0.794	0.497	0.614	0.631	0.719	OOM	OOM
HAN-3	0.594	0.685	0.673	0.616	0.603	0.687	OOM	OOM
HAN-10	0.648	0.734	0.384	0.529	0.618	0.708	OOM	OOM
HGT-3	0.799	0.850	0.733	0.797	0.888	0.916	0.837	0.878
HGT-10	0.750	0.812	0.607	0.701	0.857	0.893	0.837	0.878
HGN-3	0.746	0.809	0.814	0.859	0.903	0.927	0.815	0.861
HGN-10	0.735	0.800	0.822	0.864	0.898	0.923	0.813	0.859

at that type. We also note that some previous works [13, 25] can predict missing type information. However, they cannot be used to generate the pseudo types, as they still need partial type information from some nodes and edges as supervision.

Besides, in the third scenario, we also evaluate the heterogeneous GNNs on a complete HIN with all node/edge types given. This assesses if explicit type information is useful, and how the baselines with full type information compare to our model.

Performance on LHGs. In the first scenario, all methods do not have access to the node/edge types. We report the results in Table 2, and make the following observations.

First, our proposed LHGNN consistently outperforms all the baselines across different metrics and datasets. The results imply that LHGNN can adeptly capture latent semantics between nodes to assist link prediction, even without any type information.

Second, the performance of classic GNN baselines is consistently competitive or even slightly better than heterogeneous GNNs. This finding is not surprising—while heterogeneous GNNs can be effective on HINs, their performance heavily depends on high-quality type information which is absent from LHGs.

Third, translation models are usually worse than GNNs, possibly because they do not take advantage of node features, and lack a message-passing mechanism to fully exploit graph structures.

Performance with pseudo types. In the second scenario, we generate pseudo types for heterogeneous GNNs and translation models, and report their results in Table 3. We observe different outcomes on different kinds of baselines.

On one hand, translation models generally benefit from the use of pseudo types. Compared to Table 2 without using pseudo types, TransE- K can achieve an improvement of 13.2% and 8.5% in MAP and NDCG, respectively, while TransR- K can improve the two metrics by 5.2% and 3.6%, respectively (numbers are averaged over the four datasets). This demonstrates that even very crude type estimation (e.g., K -means clustering) is useful in capturing latent semantics between nodes. Nevertheless, our model LHGNN still outperforms translation models using pseudo types.

On the other hand, heterogeneous GNNs can only achieve marginal improvements with pseudo types, if not worse performance. A potential reason is that pseudo types are noisy, and the message-passing mechanism of GNNs can propagate local errors caused by the noises and further amplify them across the graph. In contrast, the lack of message passing in translation models make them less susceptible to noises, and the benefit of pseudo types outweighs the effect of noises.

Overall, while pseudo types can be useful to some extent, they cannot fully reveal the latent semantics between nodes due to potential noises. Moreover, we need to set a predetermined number of pseudo types, which is not required by our model LHGNN.

Performance on complete HINs. The third scenario is designed to further evaluate the importance of type information, and how LHGNN fares against baselines equipped with full type information on HINs. Specifically, we compare the performance of the heterogeneous GNNs on the two datasets DBLP and OGB-MAG, where type information is fully provided. To enhance the link prediction of the heterogeneous GNN models, we adopt a relation-aware decoder [22] to compute the score for a candidate link (x, y) as $\mathbf{h}_x^T \mathbf{W}_r \mathbf{h}_y$, where $\mathbf{W}_r \in \mathbb{R}^{d_h \times d_h}$ is a learnable matrix for each edge type $r \in R$. We report the results in Table 4.

We observe that heterogeneous GNNs with full type information consistently perform better than themselves without any type information (cf. Table 2). This is not surprising given the rich semantics expressed by explicit types. Moreover, LHGNN achieves comparable results to the heterogeneous GNNs or sometimes better results (cf. Table 2), despite LHGNN not requiring any explicit

Table 4: Evaluation of link prediction on HINs with full access to node/edge types for heterogeneous GNNs. Percentages in parenthesis indicate the improvement to their performance on LHGs (cf. Table 2).

Methods	DBLP		OGB-MAG	
	MAP	NDCG	MAP	NDCG
HAN	0.789 (+3.4%)	0.821 (+2.5%)	OOM	OOM
HGT	0.902 (+0.6%)	0.927 (+0.4%)	0.872 (+4.4%)	0.904 (+3.2%)
HGN	0.909 (+0.2%)	0.932 (+0.2%)	0.855 (+4.5%)	0.892 (+3.4%)

Table 5: Evaluation of node type classification on LHGs.

Methods	DBLP		OGB-MAG	
	MacroF	Accuracy	MacroF	Accuracy
GCN	0.376 ± 0.009	0.785 ± 0.002	0.599 ± 0.011	0.890 ± 0.003
GAT	0.310 ± 0.003	0.782 ± 0.001	0.624 ± 0.035	0.894 ± 0.007
GraphSAGE	0.477 ± 0.021	0.842 ± 0.012	0.550 ± 0.014	0.902 ± 0.004
HGT	0.464 ± 0.009	0.837 ± 0.005	0.823 ± 0.018	0.973 ± 0.003
HGN	0.292 ± 0.001	0.778 ± 0.001	0.531 ± 0.003	0.847 ± 0.003
LHGNN	0.662 ± 0.001	0.995 ± 0.001	0.884 ± 0.002	<u>0.953</u> ± 0.001

type. A potential reason is the node- and path-level semantic embeddings in LHGNN can capture latent semantics in a finer granularity, whereas the explicit types on a HIN may be coarse-grained. For example, on a typical academic graph, there are node types of Author or Paper, but no finer-grained types like Student/Faculty Author or Research/Application Paper is available.

5.3 Evaluation of Node Type Classification

To evaluate the expressiveness of semantic embeddings in capturing type information, we further use them to conduct node type classification on DBLP and OGB-MAG thanks to their accessible ground truth types. We perform stratified train/test splitting, *i.e.*, for each node type, we use 60% nodes for training and 40% for testing. For each node, we concatenate its primary node embedding and semantic embedding, and feed it into a logistic regression classifier to predict its node type. We choose five competitive baselines, and use their output node embeddings to also train a logistic regression classifier on the same split.

We employ macro-F score and accuracy as the evaluation metrics, and report the results in Table 5. We observe that LHGNN can significantly outperform the other baselines, with only one exception in accuracy on OGB-MAG. Since the node types are imbalanced (*e.g.*, authors account for 77.8% of all nodes on DBLP, and 64.4% on OGB-MAG), accuracy may be skewed by the majority class and is often not a useful indicator of predictive power. The results demonstrate the usefulness of semantic embedding in our model to effectively express type information.

5.4 Model Analyses

Ablation study. To evaluate the contribution of each module in LHGNN, we conduct an ablation study by comparing with several degenerate variants: (1) *no link encoder*: we remove the link

Table 6: Training time.

Nodes	Edges	Time	Epochs
20k	370k	1084s	24
40k	810k	1517s	11
60k	1.2M	2166s	8
80k	1.6M	2428s	6
100k	1.8M	2251s	5

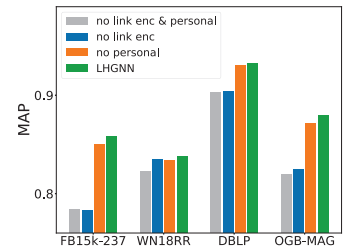


Figure 3: Ablation study.

encoder, by setting all pairwise semantic embeddings $s_{x,y}$ to zero in both training and testing; (2) *no personalization*: we remove the personalization conditioned on semantic path embeddings, by using a simple mean pooling for context aggregation; (3) *no link encoder & personalization*: we remove both modules as described earlier, which is equivalent to removing the semantic embeddings altogether.

We present the results in Fig. 3 and make the following observations. First, the performance of LHGNN drops significantly when removing the link encoder, showing their importance on LHGs. In other words, the learned latent semantics between nodes are effective for link prediction. Second, without the personalization for context aggregation, the performance also declines. This shows that the context nodes have heterogeneous relationships to the target node, and the semantic path embeddings can work as intended to personalize the contexts. Third, without both of them, the model usually achieves the worst performance.

Scalability. We sample five subgraphs from the largest dataset OGB-MAG, with sizes ranging from 20k to 100k nodes. We present the total training time and number of epochs of LHGNN on these subgraphs in Table 6. As the graph grows by 5 times, total training time to converge only increases by 2 times, since generally fewer epochs are needed for convergence on larger graphs.

Additional studies. We present results on additional model studies in Appendices E, F and G, respectively.

6 CONCLUSION

In this paper, we investigated a challenging and unexplored setting of latent heterogeneous graphs (LHG) for the task of link prediction. Existing approaches on heterogeneous graphs depend on explicit type-based information, and thus they do not work well on LHGs. To deal with the absence of types, we proposed a novel model named LHGNN for link prediction on an LHG, based on the novel idea of semantic embedding at both node and path levels, and a personalized aggregation of latent heterogeneous contexts for target nodes in a fine-grained manner. Finally, extensive experiments on four benchmark datasets show the superior performance of LHGNN.

ACKNOWLEDGMENTS

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds (Grant No. A20H6b0151).

REFERENCES

- [1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *TPAMI* 35, 8 (2013), 1798–1828.
- [2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*. 1247–1250.
- [3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*.
- [4] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *TKDE* 30, 9 (2018), 1616–1637.
- [5] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*. 135–144.
- [6] Yuan Fang, Wenqing Lin, Vincent W Zheng, Min Wu, Kevin Chen-Chuan Chang, and Xiao-Li Li. 2016. Semantic proximity search on graphs with metagraph-based learning. In *ICDE*. 277–288.
- [7] Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428* (2019).
- [8] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *CIKM*. 1797–1806.
- [9] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. MAGNN: Metapath aggregated graph neural network for heterogeneous graph embedding. In *WWW*. 2331–2341.
- [10] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*. 855–864.
- [11] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*. 1024–1034.
- [12] Huiting Hong, Hantao Guo, Yucheng Lin, Xiaoqing Yang, Zang Li, and Jieping Ye. 2020. An attention-based graph neural network for heterogeneous structural learning. In *AAAI* 4132–4139.
- [13] Jon Arne Bø Hovda, Dario Garigliotti, and Krisztian Balog. 2019. NeuType: A Simple and Effective Neural Network Approach for Predicting Missing Entity Type Information in Knowledge Bases. *arXiv preprint arXiv:1907.03007* (2019).
- [14] Binbin Hu, Yuan Fang, and Chuan Shi. 2019. Adversarial learning on heterogeneous information networks. In *KDD*. 120–129.
- [15] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*. 22118–22133.
- [16] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *WWW*. 2704–2710.
- [17] Xunqiang Jiang, Tianrui Jia, Yuan Fang, Chuan Shi, Zhe Lin, and Hui Wang. 2021. Pre-training on large-scale heterogeneous graph. In *KDD*. 756–766.
- [18] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.
- [19] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*. 2181–2187.
- [20] Zemin Liu, Yuan Fang, Chenghao Liu, and Steven C.H. Hoi. 2021. Node-wise Localization of Graph Neural Networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. 1520–1526.
- [21] Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. 2021. Tail-GNN: Tail-Node Graph Neural Networks. In *KDD*. 1109–1119.
- [22] Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. 2021. Are we really making much progress? Revisiting, benchmarking and refining heterogeneous graph neural networks. In *KDD*. 1150–1160.
- [23] George A Miller. 1995. WordNet: a lexical database for English. *CACM* 38, 11 (1995), 39–41.
- [24] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*. 3075–3081.
- [25] Arvind Neelakantan and Ming-Wei Chang. 2015. Inferring missing entity type instances for knowledge base completion: New dataset and methods. In *NAACL*. 515–525.
- [26] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*. 3942–3951.
- [27] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online learning of social representations. In *KDD*. 701–710.
- [28] Aravind Sankar, Xinyang Zhang, and Kevin Chen-Chuan Chang. 2019. MetaGNN: Metagraph neural network for semi-supervised learning in attributed heterogeneous information networks. In *ASONAM*. 137–144.
- [29] Yizhou Sun and Jiawei Han. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* 3, 2 (2012), 1–159.
- [30] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB* 4, 11 (2011), 992–1003.
- [31] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2018. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR*.
- [32] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale information network embedding. In *WWW*. 1067–1077.
- [33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.
- [34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.
- [35] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *WWW*. 2022–2032.
- [36] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI* 1112–1119.
- [37] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *TNNLS* 32, 1 (2020), 4–24.
- [38] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks?. In *ICLR*.
- [39] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. 2020. Heterogeneous network representation learning: A unified framework with survey and benchmark. *TKDE* 34, 10 (2020), 4854–4873.
- [40] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. In *NeurIPS*. 11983–11993.
- [41] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *KDD*. 793–803.
- [42] Han Zhang, Yu Hao, Xin Cao, Yixiang Fang, Won-Yong Shin, and Wei Wang. 2021. Relation prediction via graph neural network in heterogeneous information networks with missing type information. In *CIKM*. 2517–2526.
- [43] Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. In *NeurIPS*. 5171–5181.
- [44] Wentao Zhang, Yuan Fang, Zemin Liu, Min Wu, and Xinming Zhang. 2022. mg2vec: Learning Relationship-Preserving Heterogeneous Graph Representations via Metagraph Embedding. *TKDE* 34, 3 (2022), 1317–1329.
- [45] Tianyu Zhao, Cheng Yang, Yibo Li, Quan Gan, Zhenyi Wang, Fengqi Liang, Huan Zhao, Yingxia Shao, Xiao Wang, and Chuan Shi. 2022. Space4HGNN: A Novel, Modularized and Reproducible Platform to Evaluate Heterogeneous Graph Neural Network. In *SIGIR*. 2776–2789.

APPENDICES

A Algorithm and Complexity

We outline the model training for LHGNN in Algorithm 1.

Algorithm 1 MODEL TRAINING FOR LHGNN

Input: latent heterogeneous graph $G = (V, E)$, training triplets T , a set of random walk paths P_v for each node v

Output: Model parameters Θ .

- 1: initialize parameters Θ ;
- 2: **while** not converged **do**
- 3: sample a batch of triplets $T_{\text{bat}} \subset T$;
- 4: **for** each target node v in the batch T_{bat} **do**
- 5: **for** each layer $l \in \{1, 2, \dots, \ell\}$ **do**
- 6: $s_v^l \leftarrow f_s(\mathbf{h}_v^{l-1}; \theta_s^l)$;
- 7: **for** each path $p \in P_v$ that ends at context node u **do**
- 8: $s_p^l \leftarrow f_p(\{s_{v_i}^l | v_i \text{ in the path } p\})$;
- 9: $\tilde{\mathbf{h}}_{u|p}^l \leftarrow \tau(\mathbf{h}_u^{l-1}, s_p^l; \theta_\tau^l)$;
- 10: $\mathbf{h}_v^l \leftarrow \text{Aggregate}(\{\tilde{\mathbf{h}}_{u|p}^l | p \in P_v\})$;
- 11: Calculate the loss \mathcal{L} by Eqs. (10),(11) and (12);
- 12: update Θ by minimizing \mathcal{L} ;
- 13: **return** Θ .

In line 1, we initialize the model parameters. In line 3, we sample a batch of triplets from training data. In lines 4–10, we calculate layer-wise node representations on the training set. Specifically, for each node v in layer l , we calculate the semantic embeddings at the node level in line 6 and path level in line 8. Next, we personalize the contexts in line 9 and aggregate them in line 10. In lines 11-12, We compute the loss and update the parameters.

We compare the complexity of one layer and one target node in LHGNN against a standard message-passing GNN. In a standard GNN, the aggregation for one node in the l -th layer has complexity $O(\bar{d}d_h^l d_h^{l-1})$, where d_h^l is the output dimension of the l -th layer and \bar{d} is the node degree. In our model, we first need to compute the node- and path-level semantic embeddings. At the node level, the cost is $O(d_s^l d_h^{l-1})$ based on Eq. (1); at the path level, given a node with k paths of maximum length L_{max} , the cost is $O(kL_{\text{max}}d_s^l)$ based on Eq. (2). The computation of scaling and shifting vectors takes $O(kd_h^{l-1}d_s^l)$ based on Eqs. (5) and (6), and the personalization of context embeddings takes $O(kd_h^{l-1})$ based on Eq. (4). Thus, the aggregation and representation update step takes $O(kd_h^l d_h^{l-1})$ based on Eqs. (7) and (8). Furthermore, the cost to sample a path of length L_{max} is $O(L_{\text{max}})$. To sample k paths for a target node in one LHGNN layer, the overhead is $O(kL_{\text{max}})$, which is negligible compared to the aggregation cost, as L_{max} is small (less than 5 in our experiments). Therefore, the total complexity for one node in one layer is $O(kL_{\text{max}}d_s^l + kd_h^l d_h^{l-1} + kd_h^l d_s^{l-1})$. As L_{max} is a small constant and $d_s^{l-1} \ll d_h^{l-1}$, the complexity reduces to $O(kd_h^l d_h^{l-1})$. Furthermore, we can limit k , the number of sampled paths from each node, by some constant value, in which case our model has the same complexity class as a standard GNN.

B Details of Baselines

We provide detailed descriptions for the baselines.

- *Classic GNNs*: GCN [18] aggregates information for the target node by applying mean-pooling over its neighbors. GAT [34]: utilizes self-attention to assign different weights to neighbors of the target node during aggregation. Meanwhile, GraphSAGE [11]: concatenates the target node with the aggregated information from its neighbors to produce the node embedding. These models treat all nodes or edges as a uniform type and do not attempt to distinguish them.
- *Heterogeneous GNNs*: HAN [35] makes use of handcrafted meta-paths to decompose HIN into multiple homogeneous graphs, one for each metapath, then employs hierarchical attention to learn both node-level and semantic-level importance for aggregation. HGT [16] applies the transformer model, using node and edge type parameters to capture the heterogeneity. HGN [22] extends GAT by employing node type information in the calculation of attention scores. These three models require type-based information in their architectures. Given an LHG, we either assume a single node/edge type or employ pseudo types for these methods, as described in the main paper.
- *Translation models*: TransE [3] models the relation between entities as a translation in the same embedding space. TransR [19] maps entity embeddings into a relation-wise space before the translation. These models require the relation type (i.e., edge type) to be known. Similarly, we assume a single edge type or employ pseudo types for them.

C Environment

All experiments are conducted on a workstation with a 12-core CPU, 128GB RAM, and 2 RTX-A5000 GPUs. We implemented the proposed LHGNN using Pytorch 1.10 and Python 3.8 in Ubuntu-20.04.

D Model Settings

For all the approaches, we use the same output dimension as 32 for fair comparison to conduct link prediction. For the two knowledge graph datasets (i.e., FB15k-237 and WN18RR), we randomly initialize a learnable parameter vector for each entity with embedding dimension 200. We tune the margin hyperparameter α for each model in order to achieve its optimal performance. All experiments are repeated 5 times, and we report the average results with standard deviations.

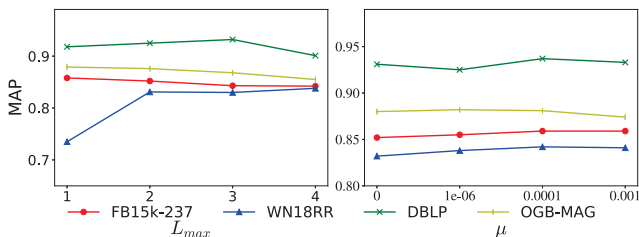
For all GNN baselines, we employ two layers with L2 Normalization on each layer. We use a margin 0.2 and dropout ratio 0.5 for all of them. For GCN, we set the hidden dimension as 32. For GAT, we use four attention heads with hidden dimension of each head as 16. For GraphSAGE, we use the mean aggregator and set its hidden dimension as 32. For all HGNN baselines, we mainly follow the default setting in [22]. In particular, we also employ 2-layer architectures for all of them. For HAN, for the node-level aggregation, we use GAT with eight attention heads with hidden dimension 8 for each head, and its dropout ratio is 0.6; and we set the dimension for semantic-level attention as 128 and set $\alpha = 1$. For HGT, we use eight attention heads, with dropout ratio as 0.2 and margin as 1. For HGN, we use eight attention heads with dropout ratio as 0.5. We set the dimension of edge embeddings as 64, and the margin as 0.5. For translation models including TransE and TransR, we

Table VII: Impact of different number of layers. In each column, the best is bolded and the runner-up is underlined.

Methods	FB15k-237		WN18RR		DBLP		OGB-MAG	
	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG
GCN-2	<u>0.790</u> ± 0.001	<u>0.842</u> ± 0.001	0.729 ± 0.002	0.794 ± 0.001	0.879 ± 0.001	0.910 ± 0.001	0.848 ± 0.001	0.886 ± 0.001
GCN-3	0.782 ± 0.001	0.837 ± 0.001	0.726 ± 0.005	0.792 ± 0.003	0.861 ± 0.001	0.896 ± 0.001	0.799 ± 0.003	0.849 ± 0.002
GCN-4	0.778 ± 0.001	0.833 ± 0.001	0.711 ± 0.005	0.781 ± 0.004	0.851 ± 0.003	0.888 ± 0.003	0.802 ± 0.002	0.851 ± 0.002
HGT-2	0.782 ± 0.001	0.837 ± 0.001	0.724 ± 0.003	0.791 ± 0.002	0.897 ± 0.001	0.923 ± 0.001	0.835 ± 0.003	0.876 ± 0.002
HGT-3	0.788 ± 0.001	0.841 ± 0.001	0.727 ± 0.012	0.792 ± 0.009	0.890 ± 0.001	0.917 ± 0.001	0.826 ± 0.002	0.870 ± 0.002
HGT-4	0.784 ± 0.005	0.838 ± 0.003	0.751 ± 0.007	0.811 ± 0.006	0.881 ± 0.011	0.911 ± 0.009	0.822 ± 0.003	0.867 ± 0.003
LHGNN	0.858 ± 0.001	0.893 ± 0.001	0.838 ± 0.003	0.877 ± 0.002	0.932 ± 0.003	0.949 ± 0.002	0.879 ± 0.001	0.909 ± 0.001

Table VIII: Impact of model size. $|\Theta|$ denotes the number of learnable parameters in a model.

Methods	$ \Theta $	DBLP		OGB-MAG		
		MAP	NDCG	MAP	NDCG	
GCN-32	12K	0.879 ± 0.001	0.910 ± 0.001	5K	0.848 ± 0.001	0.886 ± 0.001
GCN-64	25K	0.890 ± 0.001	0.918 ± 0.001	12K	0.849 ± 0.001	0.887 ± 0.001
GCN-96	41K	0.890 ± 0.001	0.918 ± 0.001	21K	0.847 ± 0.001	0.886 ± 0.001
GAT-16	26K	<u>0.913</u> ± 0.001	<u>0.936</u> ± 0.001	13K	0.830 ± 0.004	0.872 ± 0.003
GAT-32	60K	0.910 ± 0.001	0.932 ± 0.001	33K	0.828 ± 0.001	0.871 ± 0.001
GAT-48	102K	0.910 ± 0.002	0.933 ± 0.002	62K	0.815 ± 0.005	0.861 ± 0.004
HGT-32	21K	0.897 ± 0.001	0.923 ± 0.001	14K	0.835 ± 0.003	0.876 ± 0.002
HGT-64	61K	0.907 ± 0.001	0.930 ± 0.001	48K	0.836 ± 0.003	0.877 ± 0.002
HGN-32	239K	0.907 ± 0.003	0.930 ± 0.002	231K	0.818 ± 0.001	0.863 ± 0.001
HGN-64	717K	0.905 ± 0.001	0.929 ± 0.001	703K	0.826 ± 0.001	0.869 ± 0.001
LHGNN	25K	0.932 ± 0.003	0.949 ± 0.002	12K	0.879 ± 0.001	0.909 ± 0.001

**Figure IV: Impact of parameters.**

set the learning rate as 0.005, embedding dimension as 200 and the margin as 0.2.

For our model, we also employ a two-layer architecture with L2-Normalization. We randomly sample 50 paths starting from each target node. We set the dimension of semantic embeddings as 10, decay ratio λ as 0.1, the weight of scaling and shifting constraints μ as 0.0001, and margin as 0.2.

E Number of Layers

We investigate the impact of number of layers in representative baseline models. Results from Table VII show that adding more

layers might not give better results. Overall, they are still much worse than our proposed LHGNN.

F Impact of Model Size

We investigate the impact of model size (*i.e.*, the number of learnable parameters in a model) on empirical performance. Specifically, we select several representative GNNs from the baselines, and vary the size of each GNN by increasing its hidden dimension. For instance, GCN-32 indicates that its hidden layer has 32 neurons. Table VIII shows the link prediction performance of the GNNs with varying model sizes. Overall, larger models employing the same GNN architecture can only achieve slight improvements, and LHGNN continues to outperform them despite having a relatively small model. The results imply that the effectiveness of LHGNN comes from the architectural design rather than stacking with more parameters.

G Parameter Sensitivity

To study the impact of model parameters, we showcase two of the important parameters, including the maximum length for path sampling L_{max} , and the weight of scaling and shifting constraint μ . We present the results in Fig. IV. For sparse datasets with low average degree such as WN18RR and DBLP, using a large maximum length for paths can generally improve the performance, as it can exploit more contextual structures around the target node. For the other two datasets, their performance is generally less affected as the maximum length of paths increases. For the hyper-parameter μ , LHGNN generally achieves the best performance in the interval [0.0001, 0.001] across the four datasets, demonstrating the necessity of this constraint.

H Data Ethics Statement

To evaluate the efficacy of this work, we conducted experiments which only use publicly available datasets^{2,3,4}, in accordance to their usage terms and conditions if any. We further declare that no personally identifiable information was used, and no human or animal subject was involved in this research.

²<https://github.com/DeepGraphLearning/KnowledgeGraphEmbedding.git>

³https://github.com/seongjunyun/Graph_Transformer_Networks.git

⁴<https://github.com/snap-stanford/ogb.git>