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

WWW_2023

Code:None.

2023. 7. 13 • ChongQing

Reported by Yang Peng

1.Introduction

2.Method

3.Experiments

Introduction

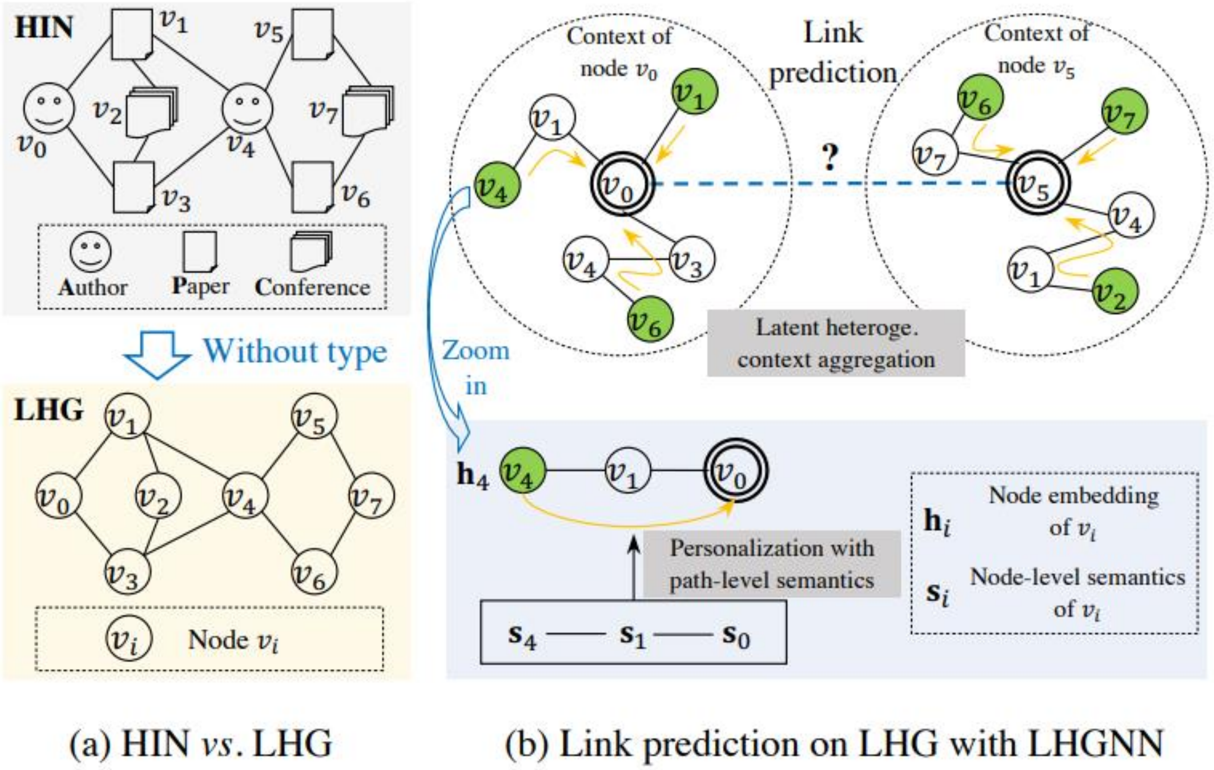


Figure 1: Illustration of our problem and approach. (a) Comparison of HIN and LHG. (b) Key insights of our approach.

Problem:
in many real-world,scenarios, type information is often **noisy, missing or inaccessible**。

Contributions:
(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.

Method

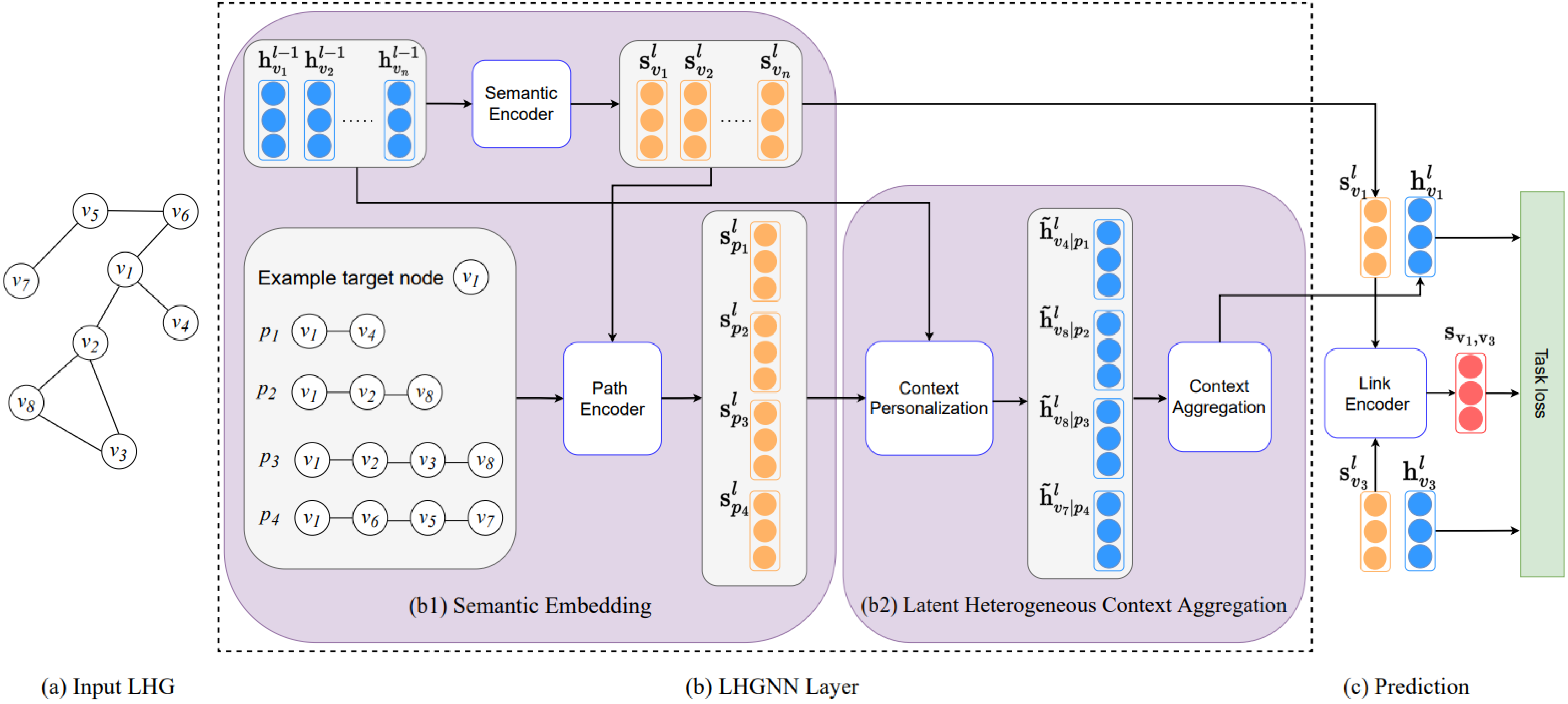


Figure 2: Overall framework of LHGNN.

Method

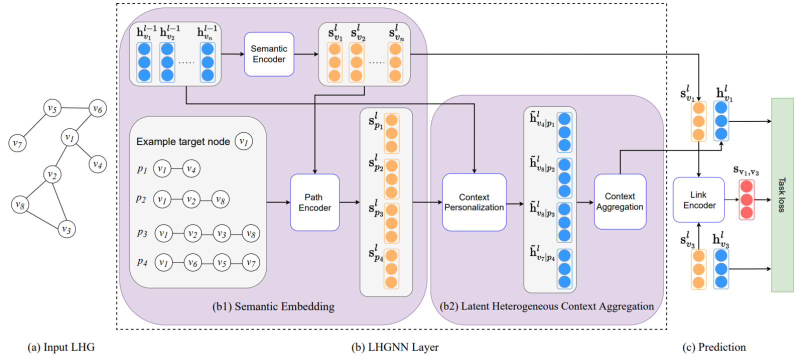
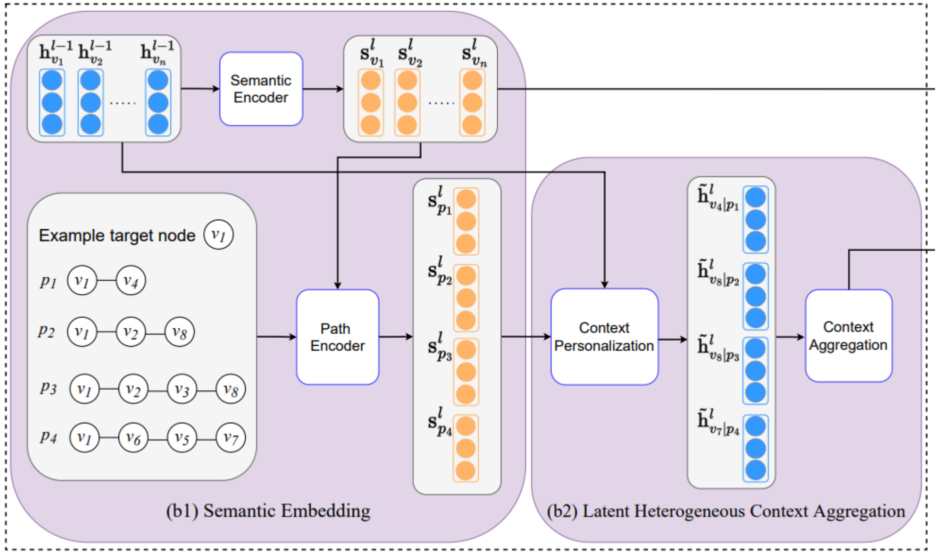


Figure 2: Overall framework of LHGNN.



Semantic Embedding

Node-level semantic embedding

primary embedding $\mathbf{h}_v$ previous layer[1] $(\mathbf{h}_{v_1}^{l-1}, \mathbf{h}_{v_2}^{l-1}, \ldots)$

$$\mathbf{s}_v^l = \mathrm{LeakyReLU}(\mathbf{W}_s^l \mathbf{h}_v^{l-1} + \mathbf{b}_s^l), \qquad (1)$$

Path-level semantic embedding

$$\mathbf{s}_{p_i}^l = f_P(\{\mathbf{s}_{v_j}^l \mid v_j \text{ in the path } p_i\}), \qquad (2)$$

path $p_i \in P_v$

P_v denote the set of sampled path

Method

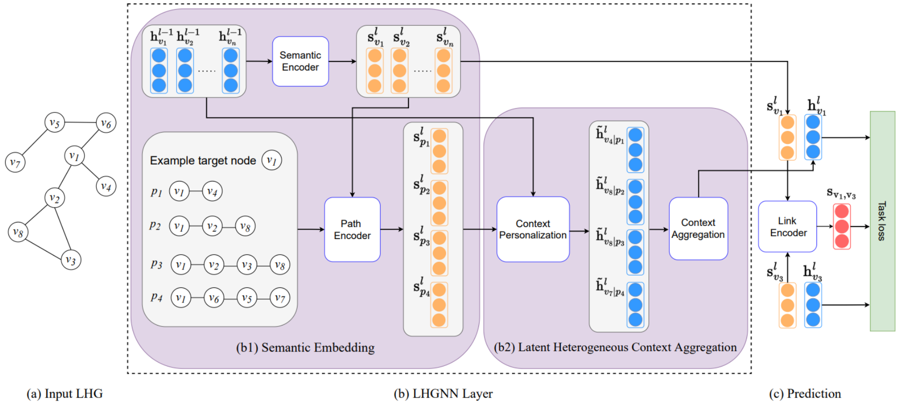
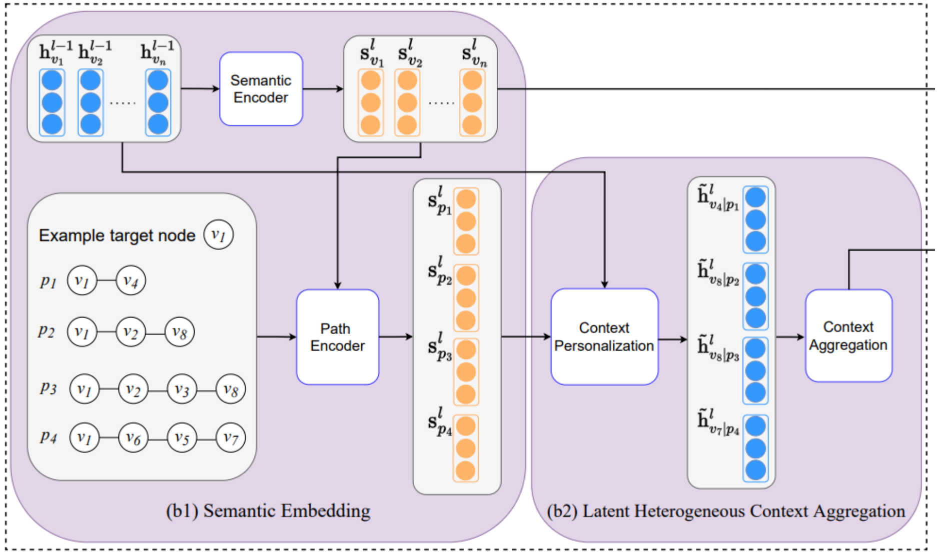


Figure 2: Overall framework of LHGNN.



Latent Heterogeneous Context Aggregation

Context personalization

$$\tilde{\mathbf{h}}^l_{u|p} = \tau(\mathbf{h}^{l-1}_u, \mathbf{s}^l_p; \theta^l_\tau), \qquad (3)$$

transformation function $\tau(\cdot; \theta^l_\tau)$ $\quad \theta^l_\tau = \{\mathbf{W}^l_\gamma, \mathbf{W}^l_\beta, \mathbf{b}^l_\gamma, \mathbf{b}^l_\beta\}.$

$$\tilde{\mathbf{h}}^l_{u|p} = (\gamma^l_p + 1) \odot \mathbf{h}^{l-1}_u + \beta^l_p, \qquad (4)$$

$$\gamma^l_p = \text{LeakyReLU}(\mathbf{W}^l_\gamma \mathbf{s}^l_p + \mathbf{b}^l_\gamma), \qquad (5)$$

$$\beta^l_p = \text{LeakyReLU}(\mathbf{W}^l_\beta \mathbf{s}^l_p + \mathbf{b}^l_\beta), \qquad (6)$$

Context aggregation

$$\mathbf{c}^l_v = \text{Mean}(\{e^{-\lambda L(p)} \tilde{\mathbf{h}}^l_{u|p} \mid p \in P_v\}), \qquad (7)$$

$L(p)$ gives the length of the path p

$$\mathbf{h}^l_v = \text{LeakyReLU}(\mathbf{W}^l_h \mathbf{c}^l_v + \mathbf{b}^l_h), \qquad (8)$$

Method

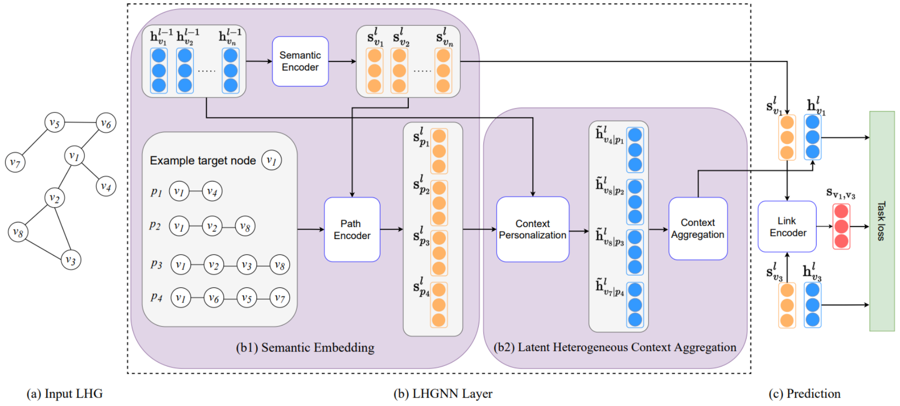
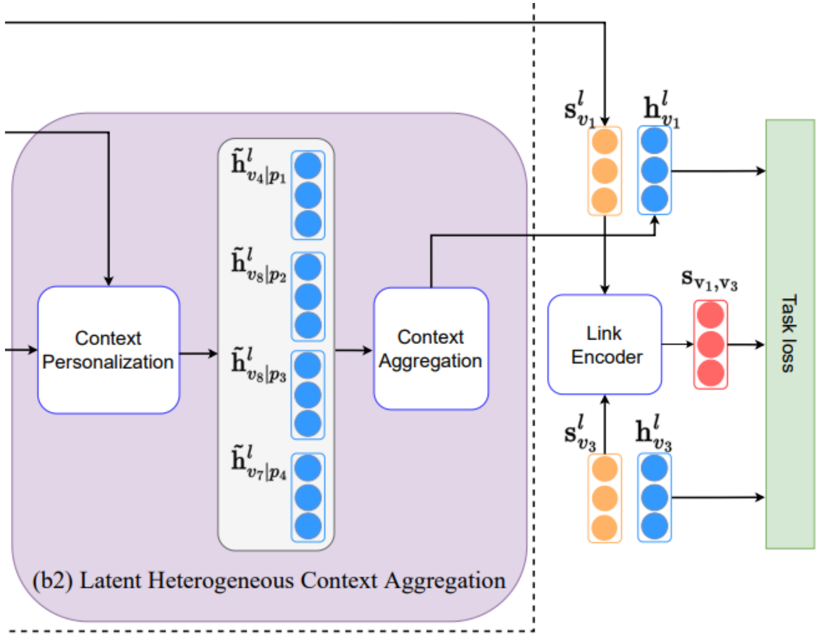


Figure 2: Overall framework of LHGNN.



Link Prediction

Link encoder

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

$$\mathbf{W}, \mathbf{U} \in \mathbb{R}^{d_s \times d_h}$$

Loss function

construct a triplet (a, b, c)

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

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

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

Experiments

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

Experiments

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

Experiments

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

Experiments

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	<u>0.477</u> ± 0.021	<u>0.842</u> ± 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

Experiments

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

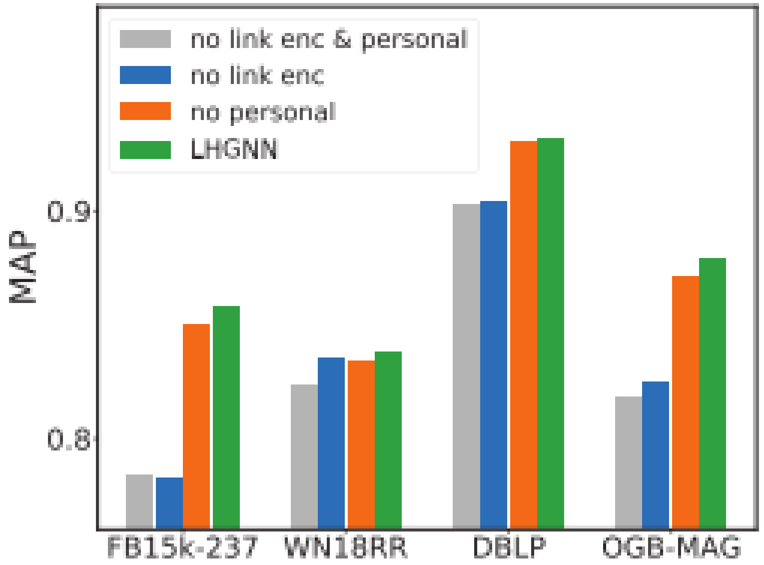


Figure 3: Ablation study.

Thank you!