Link Prediction on Latent Heterogeneous Graphs

Trung-Kien Nguyen*, Zemin Liu*, Yuan Fang

SMU Singapore Management University | School of Computing and Information Systems

NUS National University of Singapore

* Co-first authors with equal contribution.

In Proceedings of the 2023 ACM Web Conference, April 30 - May 4, 2023

Outline

- **Problem & related work**
- Proposed model: LHGNN
- Experiments
- Conclusions

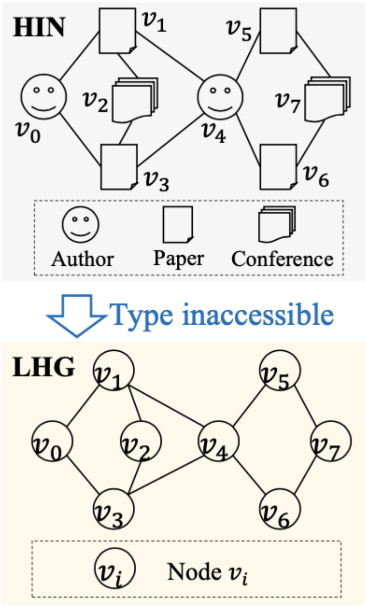
Motivation

- **Link prediction**: fundamental task on graph with many applications

- **Heterogeneous Graphs** (or Heterogeneous Information Networks/HIN) consist of multiple types of nodes/edges, providing rich and diverse semantics for link prediction.

- In real-world scenarios, type information of HINs can be noisy, missing or completely inaccessible.

  → **Latent Heterogeneous Graphs** (LHG)



(a) HIN *vs.* LHG

Related Work

- Graph Neural Networks
  - Homogeneous graphs: GCN [1], GAT [2], etc.
  - Heterogeneous graphs: HAT [3], HGT [4], etc.
- Knowledge Graph Embedding
  - TransE [5], TransR [6], etc.
- Predicting Missing Types
  - RPGNN [7], Linear.Adagrad [8], etc.

[1] Kipf, T. N., et al. 2017. Semi-supervised classification with graph convolutional networks. ICLR.
[2] Veličković, P., et al. 2018. Graph attention networks. ICLR.
[3] Wang, et al. 2019. Heterogeneous graph attention network. WWW.
[4] Hu et al. 2020. Heterogeneous graph transformer. WWW.
[5] Bordes et al. 2013. Translating embeddings for modeling multi-relational data. NeurIPS.
[6] Lin et al. 2015. Learning entity and relation embeddings for knowledge graph completion. AAAI.
[7] Zhang et al. 2021. Relation prediction via graph neural network in heterogeneous information networks with missing type information. CIKM.
[8] Neelakantan et al. 2015. Inferring missing entity type instances for knowledge base completion: New dataset and methods. NAACL.

Problem formulation

Latent Heterogenenous Graph (LHG): $(V, E, A, R, \boldsymbol{\psi}, \boldsymbol{\phi})$

- V, E: set of nodes, edges
- A, R: set of node types, and edge types, respectively
- $\boldsymbol{\psi}: V \rightarrow A$, $\boldsymbol{\phi}: E \rightarrow R$: type mapping functions
- A, R, $\boldsymbol{\psi}$, $\boldsymbol{\phi}$: inaccessible/unknown, yet still exist

Link Prediction on LHG: Given a query node, we rank other nodes by their probability of forming a link with the query.

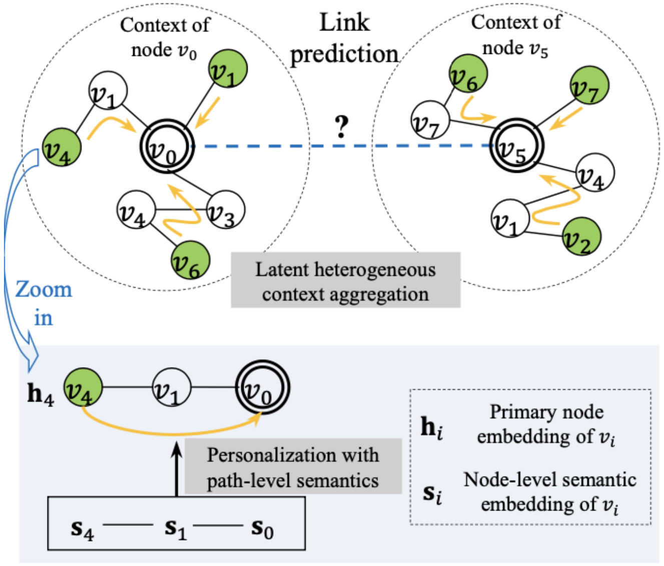
Outline

- Problem & related work
- **Proposed model: LHGNN**
- Experiments
- Conclusions

Challenges

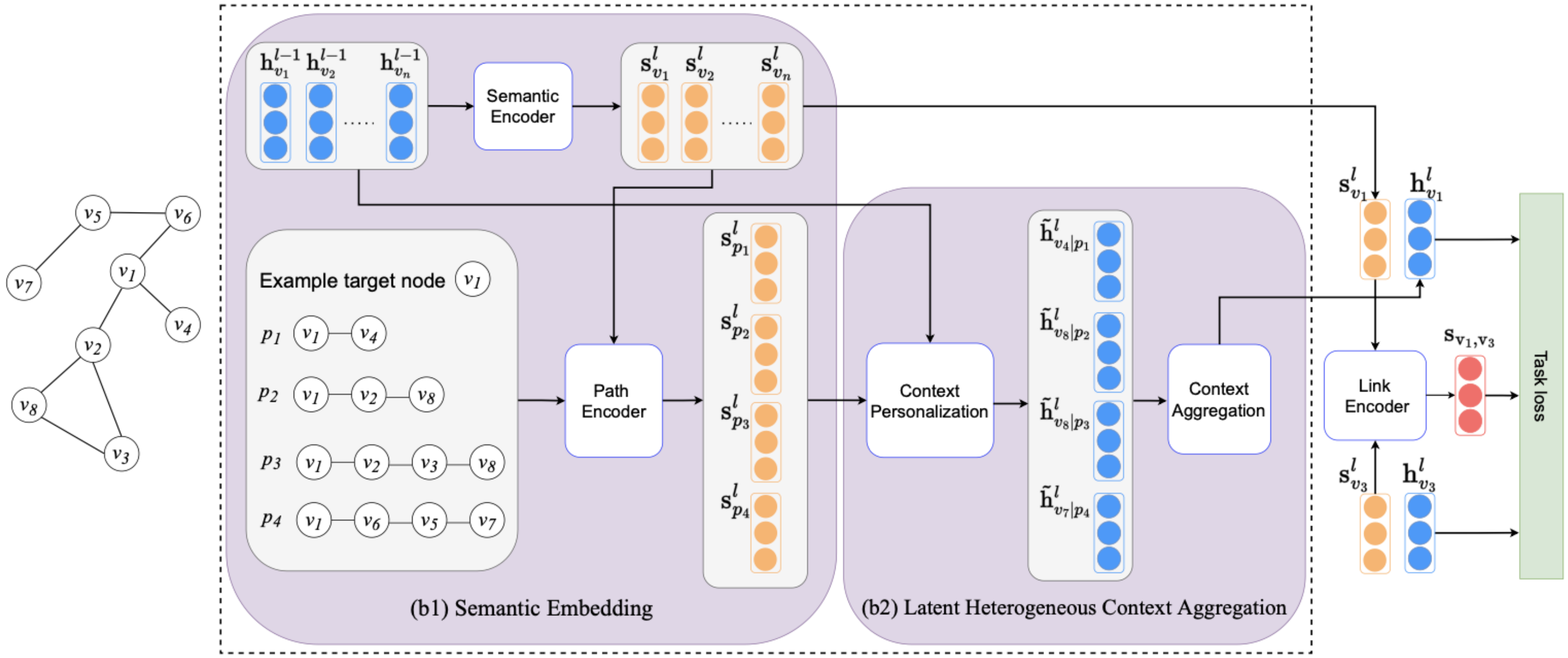
1. *How to capture the **latent semantics** on and between nodes without any type information?*

2. *As contextual nodes of a target node often carry latent heterogeneous semantics, how to **differentiate** them for finer-grained context aggregation?*



(b) Link prediction on LHG with LHGNN

LHGNN: Orverall Framework



(a) Input LHG (b) LHGNN Layer (c) Prediction

Semantic Embedding

- Node-level Semantic Embedding

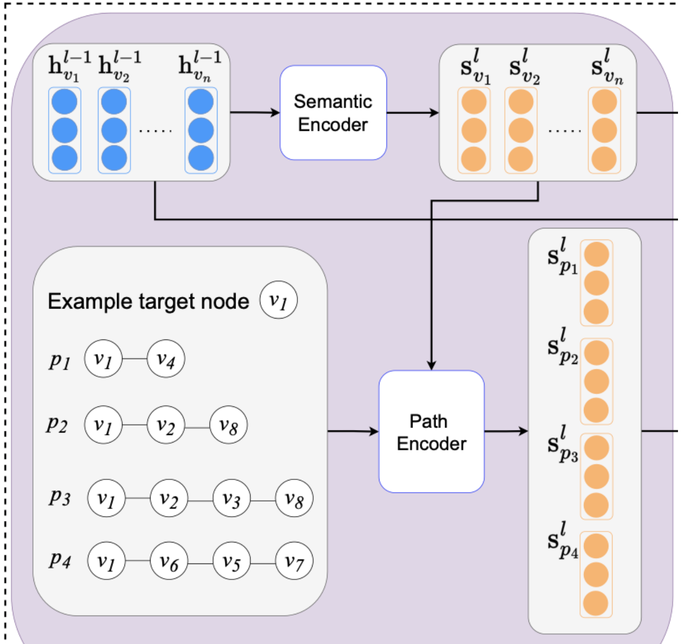
  Extracts abstract semantic information from primary node embedding

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

- Path-level Semantic Embedding

  Captures latent heterogeneous semantics from different context nodes for target node

$$\mathbf{s}^l_{p_{\langle v,u \rangle}} = f_p(\{\mathbf{s}^l_{v_i} | v_i \text{ in the path } p_{\langle v,u \rangle}\}),$$



(b1) Semantic Embedding

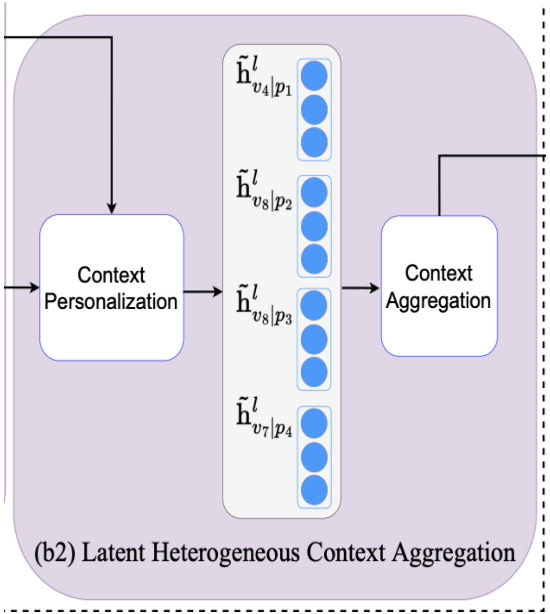
Latent Heterogeneous Context Aggregation

- Context Personalization

  Modulates latent heterogeneous semantics from different context nodes

$$\tilde{\mathbf{h}}^{l}_{u|p_{\langle v,u\rangle}} = (\gamma^{l}_{p_{\langle v,u\rangle}} + \mathbf{1}) \odot \mathbf{h}^{l-1}_{u} + \beta^{l}_{p_{\langle v,u\rangle}}$$

$$\gamma^{l}_{p_{\langle u,v\rangle}} = \text{LeakyReLU}(\mathbf{W}^{l}_{\gamma}\mathbf{s}^{l}_{p_{\langle u,v\rangle}} + \mathbf{b}^{l}_{\gamma})$$
$$\beta^{l}_{p_{\langle u,v\rangle}} = \text{LeakyReLU}(\mathbf{W}^{l}_{\beta}\mathbf{s}^{l}_{p_{\langle u,v\rangle}} + \mathbf{b}^{l}_{\beta})$$



(b2) Latent Heterogeneous Context Aggregation

Outline

- Problem & related work
- Proposed model: LHGNN
- **Experiments**
- Conclusions

Experimental setup

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
# Relations	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

Baselines

- GNN:
  GCN [1], GAT [2], SAGE [3]
- Translation:
  TransE [4], TransR [5]
- HGNN:
  HAN [6], HGT [7], HGN [8]

Metrics

- MAP, NDCG

[1] Kipf, T. N., et al. 2017. Semi-supervised classification with graph convolutional networks. ICLR.
[2] Veličković, P., et al. 2018. Graph attention networks. ICLR.
[3] Hamilton W L., et al. 2017. Inductive representation learning on large graphs. NeurIPS.
[4] Bordes et al. 2013. Translating embeddings for modeling multi-relational data. NeurIPS.
[5] Lin et al. 2015. Learning entity and relation embeddings for knowledge graph completion. AAAI.
[6] Wang, et al. 2019. Heterogeneous graph attention network. WWW.
[7] Hu et al. 2020. Heterogeneous graph transformer. WWW.
[8] Lv et al. 2021. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. KDD.

Link Prediction

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**

Node Type Classification

Table 5: Evaluation of node type classification on LHGs.

Methods	DBLP MacroF	DBLP Accuracy	OGB-MAG MacroF	OGB-MAG 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	0.953 ± 0.001

Outline

- Problem & related work
- Proposed model: LHGNN
- Experiments
- **Conclusions**

Conclusions

- **Problem**
  - Link prediction on Latent Heterogeneous Graph
- **Proposed model: LHGNN**
  - Address the absence of type information
  - Propose the novel idea of **semantic embedding** at both node and path levels to capture latent semantics
  - **Personalize the aggregation** of latent heterogeneous contexts for target nodes in a fine-grained manner
- **Experiments**

Thanks!

Paper

Code

Link Prediction on Latent Heterogeneous Graphs

Trung-Kien Nguyen*, Zemin Liu*, Yuan Fang

In Proceedings of the 2023 ACM Web Conference, April 30 - May 4, 2023