Link Prediction on Latent Heterogeneous Graphs

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

* 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)**
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.

[6] Lin et al. 2015. Learning entity and relation embeddings for knowledge graph completion. AAAI.
[8] Neelakantan et al. 2015. Inferring missing entity type instances for knowledge base completion: New dataset and methods. NAACL.
Problem formulation

Latent Heterogeneous Graph (LHG): \((V, E, A, R, \psi, \phi)\)
- \(V, E\): set of nodes, edges
- \(A, R\): set of node types, and edge types, respectively
- \(\psi: V \rightarrow A\), \(\phi: E \rightarrow R\): type mapping functions
- \(A, R, \psi, \phi\): 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?
LHGNN: Overall Framework

(a) Input LHG

(b) LHGNN Layer

(c) Prediction
Semantic Embedding

• Node-level Semantic Embedding
  Extracts abstract semantic information from primary node embedding
  \[ s^l_v = \text{LEAKYRELU}(W^l_s h^{l-1}_v + b^l_s) \]

• Path-level Semantic Embedding
  Captures latent heterogeneous semantics from different context nodes for target node
  \[ s^l_{p\langle v,u \rangle} = f_p(\{s^l_{v_i} | v_i \text{ in the path } p\langle v,u \rangle\}) \]
Latent Heterogeneous Context Aggregation

- **Context Personalization**
  
  Modulates latent heterogeneous semantics from different context nodes

\[
\tilde{h}_{u|p_{\langle v, u \rangle}}^{l} = (\gamma_{p_{\langle v, u \rangle}}^{l} + 1) \odot h_{u}^{l-1} + \beta_{p_{\langle v, u \rangle}}^{l}
\]

\[
\gamma_{p_{\langle u, v \rangle}}^{l} = \text{LEAKYRELU}(W_{\gamma}^{l}s_{p_{\langle u, v \rangle}}^{l} + b_{\gamma}^{l})
\]

\[
\beta_{p_{\langle u, v \rangle}}^{l} = \text{LEAKYRELU}(W_{\beta}^{l}s_{p_{\langle u, v \rangle}}^{l} + b_{\beta}^{l})
\]
Outline

• Problem & related work
• Proposed model: LHGNN
• Experiments
• Conclusions
Experimental setup

Datasets

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

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

Metrics
- MAP, NDCG

[5] Lin et al. 2015. Learning entity and relation embeddings for knowledge graph completion. AAAI.
# Link Prediction

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

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

## Table 5: Evaluation of node type classification on LHGs.

<table>
<thead>
<tr>
<th>Methods</th>
<th>DBLP</th>
<th>OGB-MAG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MacroF</td>
<td>Accuracy</td>
</tr>
<tr>
<td>GCN</td>
<td>0.376 ± 0.009</td>
<td>0.785 ± 0.002</td>
</tr>
<tr>
<td>GAT</td>
<td>0.310 ± 0.003</td>
<td>0.782 ± 0.001</td>
</tr>
<tr>
<td>GraphSAGE</td>
<td>0.477 ± 0.021</td>
<td>0.842 ± 0.012</td>
</tr>
<tr>
<td>HGT</td>
<td>0.464 ± 0.009</td>
<td>0.837 ± 0.005</td>
</tr>
<tr>
<td>HGN</td>
<td>0.292 ± 0.001</td>
<td>0.778 ± 0.001</td>
</tr>
<tr>
<td>LHGNN</td>
<td><strong>0.662 ± 0.001</strong></td>
<td><strong>0.995 ± 0.001</strong></td>
</tr>
</tbody>
</table>
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!

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