Task-guided Pair Embedding in Heterogeneous Network

Chanyoung Park¹, Donghyun Kim², Qi Zhu¹, Jiawei Han¹, Hwanjo Yu³

¹University of Illinois at Urbana-Champaign
²Yahoo! Research
³Pohang University of Science and Technology (POSTECH)
Network

• A ubiquitous data structure to model the relationships between entities
• Many types of data can be flexibly formulated as networks
Classical Tasks in Networks

- Node classification
  - Predict the type of a given node
- Link prediction
  - Predict whether two nodes are linked
- Community detection
  - Identify densely linked clusters of nodes
- Network similarity
  - How similar are two (sub)networks

Example: Link Prediction (Friend Recommendation)

How do we solve these network-related tasks?

→ Node embedding-based methods
Node Embedding

- Find a **low-dimensional vector representation of each node** in a graph while preserving the network structure
  - **Intuition**: Similar nodes in a graph have similar vector representations

(Deepwalk, node2vec...)

**Input**

**Output**
Related Work: Deepwalk (Perozzi et al, 2014)

- DeepWalk converts a graph into a collection of node sequences using uniform sampling (truncated random walk)
- Assuming each sequence as a sentence, they run the Skip-gram model (Mikolov et al. 2014) to learn representation for each node (like word2vec)

Random walk

\[ \mathcal{W}_{v_4} = v_4 \rightarrow v_3 \rightarrow v_1 \rightarrow v_5 \rightarrow v_1 \rightarrow v_{46} \rightarrow v_{51} \rightarrow v_{89} \]

Maximize:
\[
\begin{align*}
\Pr(v_3 | \Phi(v_1)) \\
\Pr(v_5 | \Phi(v_1))
\end{align*}
\]

Can only be applied to a network with a single type of nodes and edges.
(not to heterogeneous network)

Heterogeneous network (HetNet)

- A network with **multiple types of nodes** and **multiple types of edges**
- A lot of networks in reality are heterogeneous network

How do we embed nodes in a heterogeneous network?
Node Embedding for Heterogeneous Network: Metapath2vec (Dong et al, 2017)

- Motivation: Deepwalk assumes that each node has a single type → Extend Deepwalk to HetNet!

Task-guided HetNet embedding

• Instead of learning general node embeddings, what about we focus on a specific task?

• Example: Author Identification
  • Predict the true authors of an anonymized paper given
    • Paper abstract
    • Venue (e.g., KDD, ICDM)
    • References

• Can we predict the true authors? [1,2]

Previous Research on Task-guided HetNet Embedding

[WSDM17] Task-guided and path-augmented heterogeneous network embedding for author identification

• **Step 1:** Combine keywords, venue and references related to a paper to obtain the paper embedding

- **Supervised part:** Task-specific part

- **Unsupervised part:**

  - **Step 2:** Perform metapath2vec using embeddings learned in step 1

  \[ \text{Maximize: } \Pr(\text{\text{Document}} | \text{\text{Author}}) \]
  \[ \Pr(\text{\text{Document}} | \text{\text{Author}}) \]
Previous Research on Task-guided Embedding

[WWW18] Camel: Content-Aware and Meta-path Augmented Metric Learning for Author Identification

- Model the paper abstract using a GRU-based encoder
- Perform metapath2vec

\[ \text{Maximize: } \Pr(P_2 | A_4) \]

**Supervised part:**
- Task-specific part

**Unsupervised part:**
- metapath2vec

\[ \text{dist}(P_1, A_3) > \text{dist}(P_1, A_1) \]

Metric learning to model **direct** relationship

Skip-gram to model **indirect** relationship
Our Motivation

• Directly modeling the **pairwise relationship between two nodes** is crucial for task-guided embedding methods

• The ultimate goal is usually to model the likelihood of the pairwise relationship
  • i.e., Link probability between two nodes

• Example
  • Recommendation
    • The goal is to **model the likelihood of a user favoring an item** (i.e., user–item pairwise relationship)
  • Author identification
    • The goal is to **model the likelihood of a paper being written by an author** (i.e., paper–author pairwise relationship)

• However, previous task-guided embedding methods are **node-centric**
  Step 1. Learn task-guided **node embeddings**
  Step 2. Then, simply use inner product between two node embeddings to compute the pairwise likelihood

We devise **pair embedding** to directly model the pairwise relationship
Toy example: Author identification (Node embedding)

• Assumption
  • Bob has written multiple papers in various research areas
  • Alice only worked on “Clustering” topic

• Case 1) Node embedding
  • Should find a single optimal point to satisfy all relationship
    • Bob’s embedding: Should satisfy his relationship with various research areas
    • Alice’s embedding: Should be close to papers whose topics are “clustering”

• Question: What about a new paper on “Clustering” written by Bob?
  • It will be embedded together with “Clustering” papers, and therefore close to Alice
Our approach: Pair Embedding

• Assumption
  • Bob has written multiple papers in various research areas
  • Alice only worked on “Clustering” topic

• Case 2: Pair embedding

• Embed each paper–author pair such that each pair embedding independently captures ...
  1. Associated research topic
  2. Pair validity information
     • Whether the pair is valid or not
       = Whether the paper is written by the author within a pair

• By doing so, we want the pairs to be embedded close to each other if both of the above two conditions hold
Summary: Our goals

1. To model the **semantics** (e.g., research topic) behind the pairwise relationship

2. To model the **validity** of the pair regarding a specific task
   - This work: Author identification
     - Given a paper–author pair, whether the paper in the pair is written by the author in the pair
Proposed Method: TapEm
Overall Architecture
Proposed Method: TaPEm

• 1) Context Path-aware Pair Embedder
  • Step 1: Pair Embedder (Embedding Paper–Author Pair)
Proposed Method: TaPEm

• **1) Context Path-aware Pair Embedder**
  • Step 2: Context Path Embedder (Embedding Context Path)

Why do we consider the context path?

We can infer the research topic related to the pair \((v, u)\) by examining the path between paper \(v\) and author \(u\).

What is a context path?

A sequence of nodes between a target node pair.
Proposed Method: TaPEm

1) Context Path-aware Pair Embedder
   - Step 3: Injecting Context Information into Pairs

Objective (Pair embedding)
Predict pair using its context path
\[ P((\text{\textit{P1}}, \text{\textit{A1}}) | \text{\textit{P2}}) \rightarrow \text{\textit{A2}} \rightarrow \text{\textit{A3}}} \]

Skip Gram
\[ P(\text{\textit{P1}} | \text{\textit{A1}}), P(\text{\textit{P2}} | \text{\textit{A2}}) \]
\[ P(\text{\textit{A1}} | \text{\textit{P1}}), P(\text{\textit{A2}} | \text{\textit{P2}}) \]

\[ \mathcal{L}_{ctx}(v, u) = \sum_{c \in C_{v \rightarrow u}^P} - \log p((v, u) | c, P) \]
\[ p((v, u) | c, P) = \frac{\exp \left[ (g(v, u) \cdot f(c)) \right]}{\sum_{c' \in C_{v \rightarrow u}^P} \exp \left[ (g(v, u) \cdot f(c')) \right]} \]

Benefit
Pair embedding ≈ Embeddings of frequent context paths
→ Pair embedding encodes its related research topic
Proposed Method: TaPEm

- **2) Pair Validity Classifier** (Validity of Pair Embedding)

**Objective**
- Classify whether the pair is valid or not

\[ L_{pv}(v, u) = y_{v,u} \sigma(\pi(g(v, u))) + (1 - y_{v,u})(1 - \sigma(\pi(g(v, u)))) \]

\[ y_{v,u} = \begin{cases} 1, & \text{paper } v \text{ is written by author } u \\ 0, & \text{paper } v \text{ is not written by author } u \end{cases} \]

**Benefit**
- Enables to identify **relatively less active authors**
  - The training of the embedding is no longer solely based on the frequency (Limitation of Skip-Gram)
  - Two nodes will be embedded close to each other if
    1. Related to a similar research topic
    2. **The pair itself is valid**
Joint Objective

\[
L = \sum_{P \in S(P)} \sum_{w \in W_P} \sum_{v \in w} \sum_{u \in w}[C_v - \tau : C_v + \tau]
\]

- \( S(P) \): a set of meta-path scheme
- \( W_p \): a set of random walks guided by meta-path \( p \)
- \( \tau \): window size
- \( C_v \): position of paper \( v \) in walk \( w \)
Experiments

- Dataset: AMiner dataset
  - Extracted 10 years of data (2006 ~ 2015)
  - Removed the papers published in venues with limited publications
  - Two versions
    - AMiner-Top: Selected 18 top conferences from AI, DM, DB, IS, CV, and CL
    - AMiner-Full: All venues

<table>
<thead>
<tr>
<th>Statistics</th>
<th>AMiner-Top</th>
<th>AMiner-Full</th>
</tr>
</thead>
<tbody>
<tr>
<td># authors</td>
<td>27,920</td>
<td>536,811</td>
</tr>
<tr>
<td># papers</td>
<td>21,808</td>
<td>447,289</td>
</tr>
<tr>
<td># venues</td>
<td>18</td>
<td>389</td>
</tr>
</tbody>
</table>

**AI**: ICML, AAAI, IJCAI. **DM**: KDD, WSDM, ICDM. **DB**: SIGMOD, VLDB, ICDE. **IS**: WWW, SIGIR, CIKM. **CV**: CVPR, ICCV, ECCV. **CL**: ACL, EMNLP, NAACL
Experiments

• Baselines
  1. **Feature engineering–based** supervised method
  2. **General purpose** heterogeneous network embedding method
     • Metapath2vec [KDD17] (Dong et al, 2017)
  3. **Task-guided** heterogeneous network embedding methods
     • HNE [WSDM17] (Chen et al, 2017)
     • Camel [WWW18] (Zhang et al, 2018)
     • TaPEm_{npv} : TaPEm without pair validity classifier
Experiments: All authors (Active + Inactive)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Sup</th>
<th>MPV</th>
<th>HNE</th>
<th>Camel</th>
<th>TaPEm&lt;sub&gt;npv&lt;/sub&gt;</th>
<th>TaPEm</th>
<th>Impr.</th>
<th>Sup</th>
<th>MPV</th>
<th>HNE</th>
<th>Camel</th>
<th>TaPEm&lt;sub&gt;npv&lt;/sub&gt;</th>
<th>TaPEm</th>
<th>Impr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>T=2013</td>
<td>Rec@5</td>
<td>0.5460</td>
<td>0.5274</td>
<td>0.4874</td>
<td>0.5902</td>
<td>0.6405</td>
<td>0.6807</td>
<td>15.33%</td>
<td>0.6096</td>
<td>0.5990</td>
<td>0.6110</td>
<td>0.5458</td>
<td>0.7049</td>
<td>0.7097</td>
<td>16.15%</td>
</tr>
<tr>
<td></td>
<td>Rec@10</td>
<td>0.6227</td>
<td>0.6746</td>
<td>0.6301</td>
<td>0.7370</td>
<td>0.7677</td>
<td>0.7849</td>
<td>6.50%</td>
<td>0.6409</td>
<td>0.7317</td>
<td>0.7166</td>
<td>0.6811</td>
<td>0.8121</td>
<td>0.8237</td>
<td>12.57%</td>
</tr>
<tr>
<td></td>
<td>Prec@5</td>
<td>0.2285</td>
<td>0.2148</td>
<td>0.2051</td>
<td>0.2499</td>
<td>0.2662</td>
<td>0.2835</td>
<td>16.24%</td>
<td>0.2679</td>
<td>0.2562</td>
<td>0.2679</td>
<td>0.2393</td>
<td>0.3076</td>
<td>0.3087</td>
<td>15.23%</td>
</tr>
<tr>
<td></td>
<td>Prec@10</td>
<td>0.1323</td>
<td>0.1401</td>
<td>0.1334</td>
<td>0.1555</td>
<td>0.1632</td>
<td>0.1664</td>
<td>7.01%</td>
<td>0.1418</td>
<td>0.1595</td>
<td>0.1590</td>
<td>0.1508</td>
<td>0.1795</td>
<td>0.1818</td>
<td>13.98%</td>
</tr>
<tr>
<td></td>
<td>F1@5</td>
<td>0.3222</td>
<td>0.3052</td>
<td>0.2888</td>
<td>0.3452</td>
<td>0.3761</td>
<td>0.4003</td>
<td>15.96%</td>
<td>0.3722</td>
<td>0.3589</td>
<td>0.3724</td>
<td>0.3327</td>
<td>0.4283</td>
<td>0.4303</td>
<td>15.53%</td>
</tr>
<tr>
<td></td>
<td>F1@10</td>
<td>0.2192</td>
<td>0.2230</td>
<td>0.2202</td>
<td>0.2568</td>
<td>0.2691</td>
<td>0.2746</td>
<td>6.93%</td>
<td>0.2322</td>
<td>0.2619</td>
<td>0.2602</td>
<td>0.2470</td>
<td>0.2940</td>
<td>0.2978</td>
<td>13.71%</td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td>0.7817</td>
<td>0.8887</td>
<td>0.8614</td>
<td>0.9112</td>
<td>0.9164</td>
<td>0.9178</td>
<td>0.72%</td>
<td>0.7641</td>
<td>0.8923</td>
<td>0.8855</td>
<td>0.8768</td>
<td>0.9291</td>
<td>0.9337</td>
<td>4.64%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AMiner-All</td>
<td></td>
<td></td>
<td>AMiner-All</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T=2014</td>
<td>Rec@5</td>
<td>0.5142</td>
<td>0.5116</td>
<td>0.4665</td>
<td>0.5625</td>
<td>0.6121</td>
<td>0.6577</td>
<td>16.92%</td>
<td>0.6203</td>
<td>0.5768</td>
<td>0.5842</td>
<td>0.5494</td>
<td>0.6742</td>
<td>0.6840</td>
<td>10.27%</td>
</tr>
<tr>
<td></td>
<td>Rec@10</td>
<td>0.5792</td>
<td>0.6661</td>
<td>0.6185</td>
<td>0.7198</td>
<td>0.7471</td>
<td>0.7698</td>
<td>6.95%</td>
<td>0.6570</td>
<td>0.7114</td>
<td>0.6927</td>
<td>0.6835</td>
<td>0.7952</td>
<td>0.7998</td>
<td>12.43%</td>
</tr>
<tr>
<td></td>
<td>Prec@5</td>
<td>0.2308</td>
<td>0.2457</td>
<td>0.2284</td>
<td>0.2706</td>
<td>0.2962</td>
<td>0.3148</td>
<td>16.33%</td>
<td>0.2825</td>
<td>0.2586</td>
<td>0.2689</td>
<td>0.2529</td>
<td>0.3068</td>
<td>0.3109</td>
<td>10.05%</td>
</tr>
<tr>
<td></td>
<td>Prec@10</td>
<td>0.1447</td>
<td>0.1636</td>
<td>0.1538</td>
<td>0.1776</td>
<td>0.1851</td>
<td>0.1898</td>
<td>6.87%</td>
<td>0.1510</td>
<td>0.1623</td>
<td>0.1611</td>
<td>0.1588</td>
<td>0.1840</td>
<td>0.1850</td>
<td>13.99%</td>
</tr>
<tr>
<td></td>
<td>F1@5</td>
<td>0.3371</td>
<td>0.3320</td>
<td>0.3066</td>
<td>0.3654</td>
<td>0.3992</td>
<td>0.4258</td>
<td>16.53%</td>
<td>0.3882</td>
<td>0.3571</td>
<td>0.3683</td>
<td>0.3464</td>
<td>0.4217</td>
<td>0.4275</td>
<td>10.12%</td>
</tr>
<tr>
<td></td>
<td>F1@10</td>
<td>0.2516</td>
<td>0.2627</td>
<td>0.2463</td>
<td>0.2849</td>
<td>0.2967</td>
<td>0.3045</td>
<td>6.88%</td>
<td>0.2455</td>
<td>0.2643</td>
<td>0.2614</td>
<td>0.2577</td>
<td>0.2989</td>
<td>0.3005</td>
<td>13.70%</td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td>0.7359</td>
<td>0.8904</td>
<td>0.8619</td>
<td>0.9087</td>
<td>0.9112</td>
<td>0.9206</td>
<td>1.31%</td>
<td>0.7829</td>
<td>0.8834</td>
<td>0.8747</td>
<td>0.8770</td>
<td>0.9243</td>
<td>0.9245</td>
<td>4.65%</td>
</tr>
</tbody>
</table>

- TaPEm >> Rest (especially when N is small)
  - TapEm captures the fine-grained pairwise relationship between two nodes
    - Pushes true authors to the top ranks
- TaPEm<sub>npv</sub> > Rest
  - Pair embedding framework > Skip-gram
- TapEm > TaPEm<sub>npv</sub>
  - Pair validity classifier encodes pair validity information into the pair embedding
Experiments: Inactive Authors

• The skip-gram based model is biased to active authors
  • Most authors publish only few papers
    • 92% of authors in AMiner dataset published less than 6 publications
  • Inactive authors: Authors with less than 6 publications

<table>
<thead>
<tr>
<th>T</th>
<th>Methods</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>Camel</td>
<td>0.8854</td>
</tr>
<tr>
<td></td>
<td>TaPEm</td>
<td>0.8935</td>
</tr>
<tr>
<td>Improve.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Camel</td>
<td>0.8909</td>
</tr>
<tr>
<td></td>
<td>TaPEm</td>
<td>0.8934</td>
</tr>
<tr>
<td>Improve.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• TapEm performs much better on inactive authors
  • Benefit of pair embedding + pair validity classifier
Experiments: Case Study

- Case studies to see how TapEm ranks active authors
  - Two author groups exist within a context path
    - 1) True authors, 2) Frequently appearing false authors

- Case 1: True authors contain an active author

<table>
<thead>
<tr>
<th>Author</th>
<th>Num. publications</th>
<th>Camel</th>
<th>TapEm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiawei Han (141)</td>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Xiaolai Li (12)</td>
<td>198</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Hector Gonzalez (9)</td>
<td>296</td>
<td>81</td>
<td></td>
</tr>
</tbody>
</table>

- Case 2: Frequently appearing authors contain an active author

<table>
<thead>
<tr>
<th>Author</th>
<th>Num. publications</th>
<th>Camel</th>
<th>TapEm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yizhou Sun (23)</td>
<td>94</td>
<td>418</td>
<td></td>
</tr>
<tr>
<td>Jae-Gil Lee (10)</td>
<td>323</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td>John Paul Sondag (1)</td>
<td>1043</td>
<td>3650</td>
<td></td>
</tr>
</tbody>
</table>

- Case 3: Both author groups contain an active author

<table>
<thead>
<tr>
<th>Author</th>
<th>Num. publications</th>
<th>Camel</th>
<th>TapEm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiawei Han (141)</td>
<td>1</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Qiaozhu Mei (21)</td>
<td>4</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Dong Xin (20)</td>
<td>44</td>
<td>330</td>
<td>36</td>
</tr>
<tr>
<td>Philip S.Yu (122)</td>
<td>7</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Xiufeng Yan (36)</td>
<td>15</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Charu C. Aggarwal (30)</td>
<td>16</td>
<td>303</td>
<td></td>
</tr>
</tbody>
</table>

- Camel simply ranks active authors to high ranks (due to Skip-gram)
- TapEm is robust to the activeness of authors (due to pair embedding framework)
Experiments: Visualization of the embeddings

- **Node embedding** of TapEm
  1. More **tightly grouped together**
  2. The author embedding is closer to the cluster of the authored papers

  TaPEm generates **more accurate** embeddings for paper and author

- **Pair embedding** of TapEm
  - Makes it **even easier to distinguish** whether a pair is valid or not

  Pair embedding is useful for task-guided heterogeneous network embedding
Conclusion

• Proposed the pair embedding framework for heterogeneous network
  • Useful for tasks whose goal is to predict the likelihood of pairwise relationship between two nodes
• Directly focused on the pairwise relationship between two nodes
  • Learn the pair embedding instead of node embedding
• The pair validity classifier is effective in identifying less active true authors