HIM: Discovering Implicit Relationships in Heterogeneous Social Networks

Detailed descriptions of datasets used in the paper and complete experimental results

Bo Xu, Jinpeng Wang, Zhehuan Zhao, Hongfei Lin, and Feng Xia

1 Detailed Descriptions of Datasets

With the goal of creating a large-scale implicit relationships dataset to support research on implicit relationships discovery, data was collected from a range of sources including online social media (Weibo) and Microsoft Academic Graph (MAG). The dataset is titled HIMdata and can be downloaded from: `https://github.com/myjpgit/HIMdata.git`. Further, a public dataset on Terrorist Attacks was also used for evaluating the HIM framework. Table 1 presents an overview for the statistics of the datasets, where the symbols **#N**, **#E** and **#PR** denote the number of nodes, edges, and positive relationship (i.e. implicit relationship), respectively.

1.0.1 HIMdata

- **Same-City Relationship.** Weibo was crawled for users' information to construct a heterogeneous network based on the mutual following relationship between users. The mutual following relationship is the explicit relationship, which can be obtained by filtering according to the users' unidirectional following information. The same-city relationship

Table 1: The Datasets' Statistics

Dataset		**#N**	**#E**	**#PR**
HIMdata (Ours)	Same-City Relationship	10195	34438	16428
	Advisor-Advisee Relationship	7872	8282	2787
Public Dataset	Terrorist Attacks	1293	1648	571

here means that the geographic locations in the personal information of two users are the same city. The same-city relationship is considered as an implicit relationship.

- **Advisor-Advisee Relationship.** Microsoft Academic Graph (MAG), which contains information about scholars and publications, was used to construct an academic co-author heterogeneous network. In this dataset, the co-author relationship between two scholars is an explicit relationship. The reason for the collaboration of the two scholars may be that one scholar is the advisor of the other. Therefore, the advisor-advisee relationship is considered as an implicit relationship. The annotated details of the HIMdata will be described in Section 1.1.

1.0.2 Public Dataset

- **Terrorist Attacks**[1]. This dataset is used to build an attack-location heterogeneous network. The "co-location" relationship of terrorist attacks is regarded as an explicit relationship. The "same organization" relationship of attacks is regarded as an implicit relationship.

Table 2: Performance Comparison w.r.t. AUC and F1 with Different Training Ratios on Same-City Relationship.

Criteria	AUC			F1 score		
Tr(%)	30%	50%	70%	30%	50%	40%
GATNE	0.862	0.866	0.872	0.794	0.800	0.813
GRCN	0.817	0.828	0.832	0.676	0.706	0.709
HGT	0.847	0.858	0.862	0.766	0.776	0.779
GEN	0.808	0.813	0.822	0.672	0.680	0.698
SLICE	0.924	0.922	0.933	0.853	0.857	0.876
Shifu2	0.824	0.835	0.834	0.753	0.769	0.763
MHGCN	0.967	0.970	0.975	0.924	0.934	**0.949**
HIM(Ours)	**0.973**	**0.975**	**0.979**	**0.926**	**0.935**	0.943

1.1 Construction of HIMdata

To support the research of implicit relationships discovery, we collected user personal information from social networks (Weibo) and scholar information from Microsoft Academic Graph (MAG).

1.1.1 Same-City Relationship

To collect users' mutual following information, we crawled ∼8 million Weibo users' unidirectional following information, which was achieved by iteratively crawling the follow lists of different users. In addition, the personal information of different users (i.e. institution, certification and profile) is also collected, which is converted into user attributes with the word2vec model. For text data, we remove some irrelevant words (e.g. stop words) and non-printable characters such as emojis. Next, we eliminate users whose location information is missing and the unidirectional following information corresponding to these users, so as to determine the same-city relationship between users. By extracting repeated unidirectional following information, we successfully

Table 3: Performance Comparison w.r.t. AUC and F1 with Different Training Ratios on Advisor-Advisee Relationship.

Criteria	AUC			F1 score		
Tr(%)	30%	50%	70%	30%	50%	40%
GATNE	0.617	0.628	0.631	0.543	0.559	0.567
GRCN	0.572	0.576	0.580	0.496	0.507	0.509
HGT	0.628	0.634	0.640	0.565	0.571	0.576
GEN	0.648	0.653	0.652	0.582	0.593	0.589
SLICE	0.686	0.691	0.695	0.611	0.618	0.622
Shifu2	0.714	0.726	0.733	0.673	0.692	0.706
MHGCN	0.719	0.725	0.731	**0.685**	0.695	0.712
HIM(Ours)	**0.722**	**0.729**	**0.735**	0.683	**0.702**	**0.713**

screened out 34,438 users' mutual following relationship pairs, including 10,195 users and 16,428 same-city relationship pairs.

1.1.2 Advisor-Advisee Relationship

As mentioned previously, we use Microsoft Academic Graph (MAG) to construct a co-authors academic heterogeneous network. In order to determine the advisor-advisee relationship between scholars, we crawled 8853 records from the fields of Computer Science from The Academic Family Tree (AFT). Next, we manually eliminated any ambiguous scholar names and collated scholars' attributes such as their publications and institutions/affiliations. In addition, the related information of these publications and institutions were also collected as the heterogeneous attributes of the co-author academic network. By matching the scholar in the MAG according to scholars' names obtained from AFT, we successfully obtained 8282 co-authors relationship pairs, including 7872 scholars and 2787 advisor-advisee relationship pairs.

Table 4: Performance Comparison w.r.t. AUC and F1 with Different Training Ratios on Terrorist Attacks.

Criteria	AUC			F1 score		
Tr(%)	30%	50%	70%	30%	50%	40%
GATNE	0.919	0.926	0.931	0.829	0.835	0.847
GRCN	0.667	0.668	0.683	0.546	0.666	0.569
HGT	0.888	0.895	0.904	0.789	0.805	0.811
GEN	0.908	0.909	0.914	0.812	0.810	0.820
SLICE	0.972	0.974	0.976	0.893	0.899	0.909
Shifu2	0.853	0.860	0.862	0.793	0.811	0.818
MHGCN	0.978	0.981	0.982	0.918	0.922	0.927
HIM(Ours)	**0.984**	**0.987**	**0.988**	**0.921**	**0.929**	**0.932**

2 Experimental Results And Analysis

We compared the performance of HIM with other baseline models on three real-world datasets (Same-City Relationship, Advisor-Advisee Relationship, and Terrorist Attacks). The results are shown in Table 2, Table 3, and Table 4. The results are different when the proportion of the training set is different. As demonstrated in the results, our method, HIM, achieves the best performance on the two evaluation criteria, even when the training set is relatively small, which proves the effectiveness of our method in mining implicit relationships in heterogeneous networks. In addition, compared with GATNE, SLICE and MHGCN, the remaining network embedding methods have not achieved competitive performance in Same-City Relationship and Terrorist Attacks, because the method focusing on link prediction and the model focusing on other downstream tasks pay different attention to different information when aggregating network information.

Table 5: Performance of HIM w.r.t. AUC and F1 with Different Aggregation Attributes on Each Dataset.

Dataset	Same-City Relationship		Advisor-Advisee Relationship		Terrorist Attacks	
Criteria	AUC	F1 Score	AUC	F1 Score	AUC	F1 Score
HIM w/o *HAtt*	0.974	0.928	0.717	0.680	0.977	0.902
HIM w/o *LAtt*	0.967	0.917	**0.751**	**0.722**	0.982	0.913
HIM	**0.979**	**0.943**	0.735	0.713	**0.988**	**0.932**

2.1 Parameter Influence Analysis

We further compare the results of HIM and its variables, namely HIM w/o *HAtt* and HIM w/o *LAtt*:

- **HIM w/o *HAtt***: HIM framework without HetGNN model. In HIM-WOHET, we do not use the HetGNN model to aggregate the attributes of heterogeneous neighbor nodes, but only consider the attributes of the node itself. That is, the final attributes representation H of nodes is defined as $H = \sum_{t=1}^{T} M_t X_t W_t$, where X_t represents the initial node attributes information matrix for nodes of type t.

- **HIM w/o *LAtt***: HIM framework without link attributes weight $\lambda_{i,j}$. In HIM-WOLAW, we do not consider link attributes information in the process of aggregating network attributes information. More specifically, we directly input the adjacency matrices A_p and A_n into GCN model without multiplying link attributes weight matrix L_λ.

- **HIM**: This is the standard form of the framework as proposed in this paper. HIM uses both the HetGNN model and the link attributes weight to aggregate network attributes information.

The experimental results are shown in Table 5. When the results are considered overall, it is clear that HIM achieves the best performance. However, it is also noted that HIM does not perform better than HIM-WOHET and HIM-WOLAW on all datasets, which also confirms our previous conjecture. One of the possible reasons for this is that different information is needed to discover implicit relationships in different networks. For some networks, node attributes contribute more to the performance of the target task. The excessive extraction of network structure attributes will reduce its performance, while for other networks, the opposite can be true.

References

[1] B. Zhao, P. Sen, and L. Getoor, "Entity and relationship labeling in affiliation networks," in *ICML' 06*, vol. 4503, 2006.